# THE GENETIC BASIS OF MORPHOLOGICAL CHANGE IN CONVERGENT EVOLUTION OF NATURAL POPULATIONS

## *Identifying candidate genes behind convergent evolution in blind cave fish Astyanax mexicanus*

Martina Bradić

Dissertation presented to obtain the Ph.D degree in Biology

Instituto de Tecnologia Química e Biológica | Universidade Nova de Lisboa

Oeiras,
August, 2011

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

# THE GENETIC BASIS OF MORPHOLOGICAL CHANGE IN CONVERGENT EVOLUTION OF NATURAL POPULATIONS: IDENTIFYING CANDIDATE GENES BEHIND CONVERGENT EVOLUTION IN BLIND CAVEFISH *ASTYANAX MEXICANUS*

**Martina Bradić**

Dissertation presented to obtain the Ph.D. degree in Biology at the
Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa

Supervisors: Richard Borowsky & Henrique Teotónio

Oeiras, August 2011



*We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions. (R. A. Fisher)*

# TABLE OF CONTENTS

# List of Figures

## List of Tables

## Acknowledgements

First and foremost I want to thank to my supervisors Dr. Richard Borowsky and Dr. Henrique Teotónio. They gave me enormous support and great opportunities through all these years of my PhD. I am thankful for their excellent model as a scientist and above all as a people. I thank Dr. Borowsky for sharing his enormous knowledge about this extraordinary organism-Mexican blind cavefish as well as possibility to do the fieldwork and learn a lot about caving and caves. Henrique deserves special thanks for giving me a great opportunity to work in his lab as pre-doctoral student. He gave me enormous support through all this years of my career.

I would like to thank to my colleagues from GABBA PhD program, especially to the 10th edition. It was a great scientific and personal experience to be a student in this great environment. I would also like to thank Prof. António Amorim as a director of the GABBA 10th edition as well as our program secretary Catarina Carona for her efficacy and great help with solving any administrative problems through all these years.

I am thankful to the New York University, Biology Department for providing a great environment in the past 4 years of my research and especially to my NYU thesis committee Dr. David Fitch, Dr. Michael Purugganan and Dr. Matthew Rockman that provided a great feedback on our annual meetings. I am also very grateful for the opportunity of being a part-time lab member in Purugganan lab and especial thanks goes to Dr. Jonathan Flowers for all the scientific discussion over the lab meetings. Also, I would like to thank my current and previous lab mates at NYU; Erik Duboué and Paul Scheid as well as my colleague and friend John Burns that were there for me through all these years and provided their scientific support as well as great friendship.

I would also like to thank to our collecting expedition team; Geoffrey Hoese and Jean Luis Lacaille Muzquiz, Paco and Sarai and all the other people that provided enormous help during my fieldwork.

I want to thank all the members of Teotónio lab (Evolutionary genetics group) at the IGC that have contributed immensely not only to this thesis but also to my personal time at the IGC. They have always been a source of great mood as well as excellent scientific discussions. I thank Sara for being a great "vizinha", friend and a colleague during all this time and especially during the very intense period of writing this thesis. I would also like to thank Bruno for being a great colleague and translating my summary to Portuguese. Dr. Ivo Chelo should probably get at least 10 pages of acknowledgments for his support during my data analysis and fruitful discussions during all these years. I owe him an enormous "thank you" for being patient with my programming as well as for sharing his scripts and "R" knowledge with me.

I thank Isabel Marques and Joao Costa for sharing their experiences about genotyping and providing great technical support during my experimental work at the IGC.

I want to thank my great "IGC friend" Ricardo, for a great friendship and scientific discussions during my time at the IGC. My dearest friends Nadja and Liliana, there are no words I could possibly use to tell how much I am thankful for all these years of friendship, support and sharing the same stages of the scientific career in good and bad moments on the both sides of the ocean ☺. You guys are the greatest friends ever!

At the end, the biggest thank you goes to my parents and my sister that have been supporting me all these years wherever I was and wherever I have decided to go.

I also thank IGC for providing a great scientific environment and the Fundação para a Ciência e Tecnologia for their financial support with the grant SFRH/BD/32982/2006.

## Summary

Understanding the genetic basis of adaptive phenotypic variation is central to our understanding of the origins and maintenance of biological diversity. Repeated occurrence of the same phenotypes in closely or distantly related populations is a very powerful tool for testing the role of natural selection in maintenance of those phenotypes. Research into the molecular basis behind similar phenotypic change provides the best opportunity to unite long-standing ideas about the extent to which evolutionary change is constrained. Do similar phenotypes always diversify by the same genetic bases or does selection uses many alternative genomic routes to the same phenotypic ends? Do these changes mainly occur from already available variation in the genome or is adaptation dependent on the incoming mutation? In this dissertation we address these questions using different populations of Mexican blind cavefish (*Astyanax mexicanus*) as our model, and by taking an integrative approach using the tools of population genetics, quantitative genetics and genomics. This species is very unique, with 30 different cave populations derived from surface populations. There are numerous morphological differences between the cave adapted and closely related surface forms, including reduction in pigmentation and eye size, hypertrophy of nonoptic sensory organs, reduced metabolic rate, increased numbers of taste buds, changes in numbers of ribs as well as multiple behavioral changes. First we asked how many independent times did these morphological traits repeatedly evolved in the cave populations. We assessed genetic structure and differentiation within and among the populations using genetic data from 568 fishes from 10 cave and 11 surface localities, and 26 genetically unlinked microsatellite loci. The widespread surface localities are, with some exceptions, genetically similar to one another, whereas the cave populations are differentiated and have at least five distinct origins in the three main regions. We find lower genetic diversity in

cave populations than in related surface populations due to their smaller effective population sizes, probably because of limitations in food and space. However some of the cave populations receive migrants from the surface and exchange migrants with one another, especially when geographically close. This admixture results in significant heterozygote deficiencies at numerous loci due to Wahlund effects. In cave populations receiving migrants from the surface, we identified small numbers of individuals that are both phenotypically and genotypically intermediate between the cave and surface forms, affirming gene flow from the surface. Our study confirmed that the cave populations are derived from two main surface stocks that we call “old” and “new” populations and that diverged about 6.7 Mya, based on estimates from a previous study. “New” cave populations are closer to the surface populations while “old” cave populations are more distantly related to surface and “new” populations. In addition to that, our results suggest the old stock surface populations inhabited at least three independent cave localities while there are two independent localities inhabited by “new” stock surface populations. Thus we have established evolutionary convergence that refers to changes between “old” and “new” populations and parallel evolutionary system that refers to the changes between the populations within each of these groups. This part of the study clearly established the relationship between the phenotypically similar populations and allowed us to further investigate the importance of natural selection in the parallel and convergent evolution.

In the second part of the thesis we developed and genotyped 745 SNP markers in multiple cave and surface populations and further asked: can we find loci that were repeatedly selected for in the cave environment? All together, 80 loci were identified in several independent populations and they are potentially involved in adaptation to the cave environment.

Next, we asked where these markers are positioned in the genome and whether they coincide with regions involved in the phenotypic traits. Since the

physical genome of cavefish is not available we integrated our information with the data from laboratory crosses. We used an $F_2$ cross between the cave and surface individuals and genotyped the same SNP markers in the $F_2$ progeny. This allowed us to design a genetic map. Measures of 10 phenotypic traits that differ between cave and surface populations were available from previous studies. We used quantitative trait loci analysis (QTL) in essence correlating genotype with phenotype, to detect regions in the genome with gene loci that are responsible for each phenotype.

Some of the 80 SNPs detected as adaptive in multiple natural populations also mapped to the QTL loci for lens, amino-acid sensitivity and eye size. Those SNPs were then joined into haplotypes. Some of these haplotypes denoting putative selection were found only in “new” cave populations, but others were found both in “new” as well as “old” cave populations.

Our study supports the hypothesis that convergent adaptive phenotypic change in different populations can arise through a conserved genetic basis (shared haplotypes in new and old cave populations). Furthermore, we observed the alternative possibility that implies that natural selection can repeatedly generate similar patterns of phenotypic variation in totally novel ways (haplotypes in only new cave populations).

Finally, we asked if those selected loci represent selection/fixation of per-existing variation or new mutations. We addressed this question by comparing the ancestral allele state (surface allele) and alleles of the multiple independent populations across identified QTL regions. We observed haplotypes that were repeatedly selected in cave populations of the new lineage but were present in very low frequencies in the surface populations, or at such low frequencies as to elude detection. These suggest that adaptation from standing genetic variation plays an important role in the adaptation to the cave environment.

## Resumo

A compreensão da base genética da variação fenotípica adaptativa é central para podermos compreender a origem e a manutenção da diversidade biológica. A ocorrência sistemática dos mesmos fenótipos em populações geneticamente próximas ou distantes constitui uma técnica muito poderosa para testar o papel da seleção natural na sua manutenção. A investigação da base genética por detrás das semelhantes alterações fenotípicas constitui a melhor oportunidade de unificar ideias bem estabelecidas sobre a extensão dos constrangimentos que existem nas alterações evolutivas. Será que fenótipos semelhantes se diferenciam sob semelhantes bases genéticas ou será que a seleção usa várias vias genómicas alternativas que convergem nas mesmas soluções fenotípicas? Estas alterações ocorrem principalmente com base na variação genética já existente no genoma ou será dependente de novas mutações? Nesta dissertação abordámos estas questões de forma integrada usando diferentes populações do peixe cego das cavernas mexicano (*Astyanax mexicanus*) como modelo e tirámos partido das ferramentas da genética populacional, genética quantititativa e da genómica.

Esta espécie é única, com trinta populações diferentes em cavernas que derivaram de populações da superfície. Existem também várias características fenotípicas que diferenciam as formas das cavernas das formas da superfície, que incluem a redução pigmentar e do tamanho dos olhos, hipertrofia dos órgãos sensoriais não ópticos, taxa metabólica mais reduzida, o número de papilas gustativas e de costelas, bem como várias diferenças a nível comportamental. Em primeiro lugar perguntámos quantas vezes estas características morfológicas evoluíram de forma independente nas populações das cavernas.

Avaliámos a estrutura e a diferenciação genética entre as populações usando dados genéticos de 568 peixes de 10 cavernas e 11 locais à superfície

e 26 loci de microssatélites em segregação independente. As populações de grande parte dos locais à superfície são geneticamente semelhantes, com algumas excepções, enquanto as populações das cavernas estão geneticamente diferenciadas e têm pelo menos cinco origens nas três regiões principais. Encontrámos uma menor diversidade genética nas populações das cavernas relativamente às populações da superfície relacionadas devido ao menor tamanho efectivo das primeiras, o que por sua vez se pode justificar pelas limitações de alimento e de território. Contudo, algumas das populações das cavernas receberam migrantes da superfície e trocaram também migrantes entre si, sobretudo quando geograficamente próximas. Esta mistura resulta em deficiências significativas de heterozigotas em numerosos *loci* devido ao efeito de Wahlund. Nas populações das cavernas que receberam migrantes da superfície, identificámos pequenos números de indivíduos com fenótipos e genótipos intermédios, o que confirma o fluxo genético a partir da superfície. O nosso estudo confirmou que as populações das cavernas são derivadas de dois principais grupos a que chamamos “antigas” e “recentes” populações e que divergiram há cerca de 6.7 milhões de anos, com base nas estimativas do estudo anterior. As populações “recentes” estão mais próximas das populações da superfície, enquanto as populações “antigas” têm uma relação mais distante quer com as populações “recentes”, quer com as populações da superfície. Para além disto, os nossos resultados sugerem que as populações do grupo antigo habitaram pelo menos três locais independentes enquanto as populações do grupo mais recente habitaram dois locais independentes. Estabelecemos, deste modo, um sistema de evolução convergente que se refere a alterações entre populações “antigas” e “recentes” e um sistema de evolução paralela que refere alterações entre as populações dentro de cada um destes grupos. Esta parte do estudo estabeleceu claramente a relação entre as populações fenotipicamente semelhantes e permitiu-nos observar com maior detalhe a importância da

seleção natural na evolução paralela e convergente.

Na segunda parte da tese desenvolvemos e genotipámos 745 marcadores de SNPs em múltiplas populações de superfície e de caverna e perguntámos: podemos encontrar os *loci* que foram repetidamente selecionados no ambiente cavernícola? Em suma, 80 *loci* foram identificados em várias populações independentes e estão potencialmente envolvidos na adaptação a este ambiente.

Seguidamente, perguntámos onde se posicionavam estes marcadores no genoma e se coincidiam com a região envolvida nas características fenotípicas. Uma vez que o genoma físico do peixe cego das cavernas mexicano não está disponível, integrámos a nossa informação com dados de cruzamentos no laboratório. Usámos a geração $F_2$ de cruzamentos entre indivíduos das cavernas e da superfície e genotipámo-la, procedimento que nos permitiu obter um mapa genético. As medições de 10 fenótipos que diferenciam estas populações estavam disponíveis através de um estudo anterior. Utilizámos posteriormente uma análise de *loci* de caracteres quantitativos (QTL), essencialmente para correlacionar genótipos e fenótipos e detectar os *loci* no genoma responsáveis por cada fenótipo.

Alguns destes 80 SNPs detectados como adaptativos em várias populações naturais também foram mapeados pela análise QTL no cristalino, na sensibilidade a aminoácidos e no tamanho dos olhos. Estes SNPs foram posteriormente concatenados em haplótipos. Em alguns destes, identificámos uma indícios de seleção apenas nas populações “recentes”, mas outros haplótipos foram encontrados em ambas as populações “recentes” e “antigas”. O nosso estudo favorece a hipótese de que a semelhantes alterações fenotípicas podem surgir a partir de com base genética conservada (haplótipos partilhados nas populações das cavernas recentes e antigas). Além disto, observámos também que os fenótipos podem surgir através de alterações não conservadas (haplótipos apenas nas populações das cavernas recentes).

Por último perguntámos, os *loci* selecionados são o produto de fixações de variação preexistente ou do surgimento de novas mutações? Abordámos esta questão comparando o estado alélico ancestral (alelo da superfície) com o das várias populações independentes ao nível dos QTL identificados. Verificámos a existência de haplótipos que foram repetidamente selecionados nas populações das cavernas da nova linhagem e que estavam presentes em frequências muito baixas ou indetectáveis nas populações da superfície.

Estes dados sugerem que a adaptação a partir de variação genética preexistente tem um papel importante na adaptação ao ambiente cavernícola.

# CHAPTER 1

## INTRODUCTION

A suite of structural, functional and behavioral changes of the organism generally accompanies adaptation to a new environment and these processes have been the subjects of scientific inquiry for a long time. However the mechanism of these changes in the natural populations remain largely unknown: What are the “loci of adaptation” responsible for the emerging of the new morphological traits? How many loci are responsible for a particular trait of interest and how repeatable is evolution if the same morphology evolves multiple times from the same or divergent ancestors? What are the underlying evolutionary mechanisms that drive these changes?

Research into the molecular basis behind convergent phenotype provides the best opportunity to unite long-standing ideas about the extent to which evolutionary change is constrained, with ideas about the architecture of adaptive differences within and between populations. These are fundamental issues when considering the origins of adaptive variation and the generation of biodiversity. In this respect, the adaptation to the cave environment possesses a combination of attributes that make it a particularly powerful system for gaining a better understanding of how similar phenotypes are produced and a fuller appreciation of the origins, maintenance, and modification of diversity.

We addressed these questions here using different populations of Mexican blind cavefish (*Astyanax mexicanus*) as our model, and by taking an integrative approach using the tools of population genetics, quantitative genetics and genomics. Cavefish, widely distributed in the caves of North-East Mexico, provides a very suitable system for the study of local adaptation due to the morphological evolutionary convergences of multiple populations. This system offers a unique opportunity to investigate whether evolution of similar

phenotypes occur thorough changes in the same or different genetic loci. Due to the good surrogate for the ancestral phenotypic state (surface fish) there is also a good reason to ask if the adaptations to the novel environment are the result of new mutations or preexisting genetic variation in an ancestral population.

This thesis addresses the above-mentioned questions and focuses on assessing the relative contributions of different evolutionary forces; gene flow, and natural selection and different source of variation; new mutations and preexisting genetic variation, to the evolution of similar phenotypes in independent populations.

## 1.1. Convergence and Parallelism: similar or different ways to the same trait?

*Definitions*

Evolutionary change frequently follows a common pathway because of similar environmental pressures. It culminates in similar morphological organization, even though the plants and animals that follow such similar paths may be unrelated or only distantly related. These multiple origins of a trait represent exceptional replicates of evolutionary processes and can provide extremely valuable insights into the constraints and opportunities that govern evolution. Phenotypic similarity can occur in all levels of taxa ranging from microbes to plants and primates. Taxonomists historically classified this phenomenon and it is divided in two categories, parallelism and convergence. Parallel evolution was defined as independent occurrences of similar changes in groups with a common ancestry [1, 2]. For example two morphotypes of stickleback fish; one with reduction of pelvic structures and one with normal pelvis depending on the marine or freshwater habitat is a good example of parallelism (details discussed later) [3-10]. In contrast, similar phenotype that occurs separately in two or more lineages without a common ancestry is determined as

convergence [1] (pp. 78–79) [11, 12] (e.g. wings in birds and bats).

However, when focusing on two biological levels- phenotype and genotype- within the simplified framework of parallel and convergent changes these definitions are much more complex. There has been lot of debate in the field in order to establish common terminology for convergence and parallelism taking both phenotypic and genotypic observations into account [13, 14]. Parallel evolution is often difficult to differentiate from convergence and some authors have even suggested a continuum between convergent and parallel evolution [13, 15, 16]. The distinction between convergent, parallel, and divergent evolution indeed requires the historical evolutionary aspect of studied lineages.

One of the possible views on the parallelism *vs.* convergence on the molecular level is shown in the figure 1.1. Parallelism in this case refers to the independent evolution of the same derived state from a common ancestral state (the two Gs from T, or the two Gs from C). In contrast, convergence involves the evolution of the same derived state from different ancestral states (Figure 1.1.) [2]. These definitions were also further used in this study.

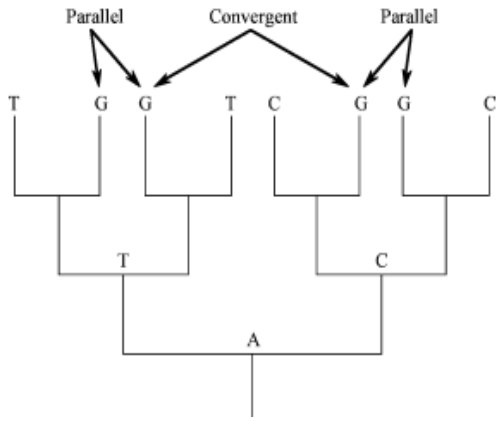



Figure 1.1. Parallelism *vs.* convergence in molecular evolution. Character states at a single nucleotide site are mapped onto a gene tree. Parallelism refers to the independent evolution of the same derived state from a common ancestral state (the two Gs from T, or the two Gs from C). In contrast, convergence involves the evolution of the same derived state from different ancestral states (G derived independently from T and C) (adapted from [2]).

*How did unrelated species or populations evolve to look so similar?*
*Evidences from experimental evolution*

Repeated patterns of phenotypic traits are commonly regarded as evidence of adaptation under common selection pressures such as common environmental factors [5, 17-23]. Despite the scientific profundity of this question, as well as the exceptional utility of convergent and parallel evolution for teaching us about adaptation and natural selection, relatively little is known about the genetics behind phenotypic similarity.

Probably among the clearest examples of parallelism and convergence on the genetic level come from experimental evolution studies. These approaches are mostly applied to organisms with short generation times and the possibility to easily identify genetic variation [2, 13, 20, 24-26]. The major advantage of these experiments is that one can control both the selective pressures as well as the evolutionary history of those populations. For example, Bull et al. used a bacteriophage (a virus that attacks bacteria) experimental evolutionary system

to examine the extent and dynamics of molecular changes during adaptation [27]. Replicate lineages were adapted to growth at high temperature on either of two bacterial hosts. The researchers then documented the extent to which convergent evolutionary changes occurred during this period. They found that more than half of the 119 observed nucleotide substitutions were the same. Some of these molecular changes were host-specific, and others were found in phages growing on both hosts. There are more evidences that phenotypic shifts in such a simple organisms result frequently in minor sequence changes and a single locus accounts for the entire response to selection [25, 28] (Table 1.1 A).

However, already in a slightly complex organism the relationship is not so clear. For example, Cooper et al. derived twelve lines of *E. coli* from a single ancestral clone. These lines evolved for 20,000 generations under low glucose medium and showed both similar and different genetic changes under the same environment. This study shows that even when genetically identical replicates are exposed to identical selection the same derived phenotype can be attained via different genetic pathways [29] (Table 1.1 A). It seems from these studies that the dynamics of selection in simple systems (like bacteriophage) might not be representative of more complex organisms, as shown in bacteria. Thus, in higher taxa with larger and more complex genomes we might expect selective constrains due to genetic background or antagonistic pleiotropy (discussed later) [2].

| Organism | Character Comparison | Method |
|---|---|---|
| Virus | Novel host | 3 |
| (FX174) | Novel host, temperature | 4 |
| Bacteria | Glucose limited media | 3 |
| (*Escherichia coli*) | Novel carbon source | 3 |
| | Thermal adaptation | 4 |
| Fungus | Carbon source | 3 |
| (*Saccharmoyces*) | | |
| Fruit fly | Wing vein | 1 |
| (*Drosophila spp.*) | Knockdown resistance | 1 |
| | Learning | 1 |
| Atlantic salmon | Domestication | 3 |
| (*Salmon salar*) | | |
| Mexican cavefish | Pigment loss 1 | 1 |
| (*Astyanax spp.*) | Eye loss 1 | 1 |
| Whitefish | Body size | 3 |
| (*Coregonus lavaretus*) | | |
| Threespine stickleback | Lateral plate reduction | 1,4 |
| (*Gasterosteus aculeatus*) | Pelvic reduction | 5 |
| Domestic mouse | Nest building | 1 |
| (*Mus domesticus*) | | |
| Rock pocketmouse | Pigment gain | 4 |
| (*Chaetodipus intermedius*) | | |
| Beach mouse | Pigment reduction | 4 |
| (*Peromyscu spolionotus*) | | |

Method of Comparison: 1 = hybrid complementation; 3 = patterns of gene expression; 4 = sequencing of candidate genes; 5 = phenotypic comparison.

Table 1.1.A. Examples in which similar phenotype evolved within a species by different genetic changes. Adapted from [13].

| Gene | Organism | Comparison |
|---|---|---|
| Mc1r | Pocket mice | 4 |
| (Pigmentation) | Several felids | 4 |
| | Little striped whiptail lizard | 4 |
| | Lesser earless lizard | 4 |
| | Snow goose | 4 |
| | Arctic skua | 4 |
| | Beach mice | 4 |
| | Mammoth | 4 |
| Opsin | Various birds | 4 |
| (UV color vision) | | |
| Pitx1 | Threespine stickleback | 3 |
| (Pelvic reduction) | Ninespinestickleback | 3,5 |
| Manatee | | 5 |
| Lysozymes | Leaf monkeys | 4 |
| (Digestive enzyme) | | |
| Ion channels | Drosophila melanogaster | 4 |
| | Homo sapiens | |
| Knox-Arp | Lycophytes and | 4 |
| (Leaf formation) | euphylophytes | |

Method of Comparison: 1 = hybrid complementation,3 = patterns of gene expression; 4 = sequencing of candidate genes; 5 = phenotypic comparison.

Table 1.1.B. Examples in which similar phenotype evolved within a species by similar genetic changes. Adapted from [13].

*Evidences from natural populations*

The genes that are the direct targets of selection for producing similar phenotypes have been identified in only a handful of natural systems. Some of the best examples in natural populations come from different fish populations with clearly established phylogenetic relatedness between different populations. For example, two eco-morphs of the lake whitefish (*Coregonus clupeaformis*), one dwarf (smaller, more limnetic) and other normal (benthic), are found in northeastern North America. They have evolved rapidly, independently and in parallel in different freshwater lakes [30-37]. Parallel

expression patterns have been identified and strongly suggest a role of natural selection in genes that are related to energetics. Another example represents genotype-phenotype association of the vision genes in cichlids. Mutation and expression in opsin genes has demonstrated parallel adaptation to a different water depth by divergent selection in Lake Victoria [38].

Finally, the most studied example comes from sticklebacks, which are interesting fishes in that they lack scales and instead have armour plating. From a highly plated marine ancestor, in numerous freshwater environments armour plating is reduced or lost repeatedly and independently after colonization. Most freshwater populations have low-plate *Eda* alleles arguing for a strong role of repeated, independent positive selection in freshwater environments from standing genetic variation [4, 39, 40]. Recent genome-wide study of adaptive loci sticklebacks has also shown *Eda,* as well as multiple other regions involved in adaptive process to freshwater environment [5, 31].

Besides the parallel genetic changes within the species those conserved changes are also apparent when comparing different species. An example of that is the adaptation of the color in the beach mice. The change in coloration is due to the exact same amino acid polymorphism in Mc1r gene that is also found in a population of wooly mammoths (Table 1.1B). Surprisingly, in the beach mice there is also a possibility that different genes from that melanin production pathway can produce the same phenotypic change (Table 1.1A and B) [4-6, 9, 10, 41]. Also, an interesting example of convergent evolution of the wing color switch genes in two different butterfly species was also described recently [42-46].

As summarized in the Table 1.1 A and B and the above mentioned examples, in some studies of the genetic basis of phenotypic similarity among closely related populations the underlying genetic mechanism is the same [2]. Because of that some have argued that parallel evolution is caused by genetic constraints; similar phenotypes evolve in parallel because genetic and

developmental constraints cause limitations to a few alternative phenotypes [2, 8, 13, 24, 45, 47]. Other researchers provide evidences of attaining the same phenotype via different genetic changes even among closely related taxa (Table 1.1A). These studies suggest that if multiple developmental pathways can lead to the same phenotype, then parallel evolution is a signal of adaptation [48]. Most of these studies are based on little empirical data, since the genome wide screen of adaptive variation in natural populations was not available till recently. Thus, it remains unclear to what extent natural selection or genetic drift can facilitate parallelism and convergence on genome-wide level. Natural selection will be at least partially responsible for repeated evolution of the same trait in association with the similar environmental change [5, 17, 49-51]. Genetic drift will also play a role in repeated phenotypic evolution, but the phenotypic transitions will probably not be linked to the environment consistently. Because of that, occurrence of the repeated phenotypes is one of the best tools to test natural selection in the wild [18, 50, 52].

In the most simplified case (one gene-one phenotype) each phenotype in nature can be a product of different patterns such as: mutations in the different genes, same genes but different position and in the most conserved example the same gene and the same nucleotide (reviewed in [53]). Any of these changes could be a result of already available variation (standing genetic variation) or they can come from new mutations. It is predicted that the adaptation from the standing genetic variation would be the fastest mode of adaptation to the new environment, since the variation is already available and it has been segregating in the same genetic background for a long time. On the other hand, it takes time for the new beneficial mutation to occur and this process would lead to much slower adaptation [8, 31, 45, 54-60]. These observations raise the following questions: do the closely related populations/species evolve their phenotype through the same genetic change-

from standing genetic variation? What do distantly related populations/species tell us about that processes? Is there any rule?

There are not many studies that address the importance of standing genetic variation *vs.* new mutations even in model species (reviewed in [53]). The reason for the low number of studies is probably because standing genetic variation is most reliably distinguished from *de novo* mutation by sampling of the ancestral and derived population. Thus, a suitable natural system that affords access to both ancestral and derived states is required (reviewed in [59]). Also, until recently it was impossible to sample enough polymorphism in non-model organisms to allow for this kind of study.

Many studies, either in the lab or in nature, point out a huge importance of standing genetic variation in the process of adaptation to the new environments. For example, adaptation from standing genetic variation in replicated populations in experimental evolution in *Drosophila* has shown that adaptation to a new laboratory environment largely occurs from the sorting and recombination of standing genetic variation at multiple loci [58]. In natural populations only one study tests the allele frequency change from standing genetic variation tracking *Eda* genotype frequencies over a multiple generation in stickleback populations. The low frequency beneficial allele present in ancestral population increased in frequency very fast over multiple generations [39, 54, 55, 59]. Recent genome wide studies or sticklebacks and whitefish also suggest that most of the adaptive variation is present in a very low frequency in the ancestral population [5, 31].

### 1.2. *Astyanax Mexicanus* as a model to test convergent and parallel evolution

Each individual cave is a single evolutionary experiment, which is replicated multiple times across the same environmental cue when additional caves are added to the study. Furthermore, the attribute of having ancestral and derived

morphological form (cave and surface) sets an excellent opportunity to study the genetic and evolutionary mechanisms of convergent and parallel evolution. Repeated appearance of the same phenotype in the group of cave organisms is also very common, especially following the repeated multiple independent colonization of caves by the surface ancestral phenotypic form which frequently result in eye regression like in cave amphipods [61, 62] and cavefish [63-67]. Organisms existing in such circumstances often result in a suite of changes called troglomorphy; progressive elongation of body form and appendages as well as an increase in sensory structures, hypertrophy of nonoptic sensory organs and a reduced metabolic rate [68-70]. Cave animals represent one of the best examples of convergent evolution and offer some unique advantages for studying its mechanisms.

One of the best-studied taxa is the teleost *Astyanax mexicanus*, a fish species that includes both eyed surface and eyeless cave-dwelling populations [64, 65, 67]. The first Mexican cave characin was described in 1936 by Hubbs and Innes as *Anoptichthys jordan*i. In the mid-1960's, as a result of activities by members of the Texas-based Association for Mexican Cave Studies in the Sierra de El Abra many different localities with the cavefish populations have been discovered. Cavefish was first described as three species (*Anoptichthys jordani, A. antrobius, and A. hubbsi*, respectively). Nowadays, we are taking about unique genus with the inter-fertile surface and cave forms and they are considered as morphotypes of the same species, *Astyanax mexicanus* [63, 65]. Multiple trips and studies of Sierra de El Abra followed and today we know of 29 cave populations of Mexican blind cavefish [65]. One of the pioneering studies to address the origin of different cavefish populations was a cross between two geographically isolated cavefish populations that resulted in $F_1$ progeny with a greater degree of eye development than exhibited by either parent indicating that mutations in different genes are involved in eye regression [66]. Molecular work followed this observation and electrophoretic

study showing minimal divergence in 17 allozyme loci concluded that the Sierra de El Abra cavefish had a common origin [63]. In contrast, Mitchell et al. who surveyed 29 different cavefish populations in the Sierra de El Abra, Sierra de Guatemala, and Micos region, proposed several different origins of Astyanax cavefish [65]. Cavefish is thus an attractive model to study genetic basis of independently evolved morphological traits.

*Evolutionary history*

Previous phylogeographic studies of *Astyanax* cavefish, using microsatellite and mtDNA, showed that the cave populations are derived from at least two different surface stocks that inhabited the Sierra de El Abra and nearby regions in succession [71-74]. The estimates from mtDNA suggest that these two groups diverged about 6.7 Mya [72]. Surface forms of the older stock originally inhabited the rivers in the El Abra region and were the likely ancestors of a series of cave populations: La Cueva de El Pachón, La Cueva de los Sabinos, El Sótano de la Tinaja, La Cueva de la Curva, El Sótano de Yerbaniz, El Sótano de Las Piedras, and La Cueva Chica (Figure 1.3). Subsequently, the surface fish of the old stock went locally extinct. The region was then invaded by another stock of *A. mexicanus*; its descendants are the current occupants of the region's surface waters and a second set of cave populations: La Cueva del Río Subterráneo, and El Sótano de Molino and Cabalo Moro (Figure 1.3). These studies indicate that independent cave populations have evolved different mechanisms of degeneration [63, 67, 75-78]. However, a sufficient detailed understanding of the biogeography, population structure and gene flow between, as well as number of independent phenotypic adaptation to the cave environment remains largely undetermined.

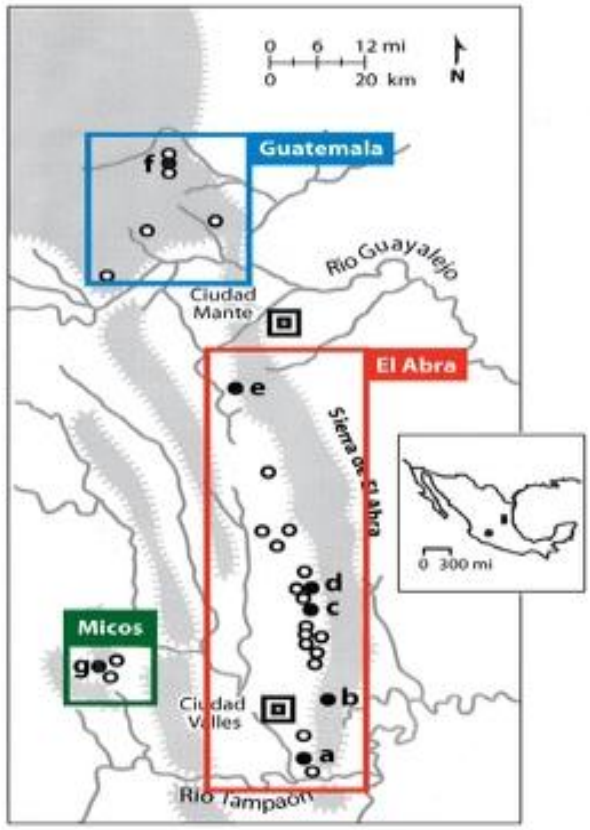



Figure 1.2. Map showing the region containing 29 different Astyanax cavefish populations in northeastern Mexico. The spheres indicate the approximate position of caves with Astyanax cavefish. The Guatemala, El Abra, and Micos clusters are indicated on the map. Inset: Mexico showing the northeastern region indicated in the sketch map (shaded rectangle) and the outlying Guerrero population (shaded sphere). Adapted from Jeffery [79]. Annu.Rev.Genet.43: 25-47.

*Morphological changes in Astyanax mexicanus*

The two forms of *Astyanax* have favorable attributes, including descent from a common ancestor, easy laboratory breeding, and the ability to perform genetic analysis, permitting their use as a model system to explore questions in both evolutionary biology and development. Although morphological changes in cavefish are similar between different populations but cave environment and

the morphs themselves also vary to a large degree. For example, the caves have different amounts of water and numbers of pools; therefore the size of the population will vary significantly. Caves can also be connected with the surface rivers or completely isolated, with consequences for the degree of troglomorphy attained by the populations, especially eye and pigmentation reduction [65, 67]. Phenotypic similarity between different cave morphs might mask mechanistic or developmental differences, making the classification of phenotypic evolution dependent on the level of organization being studied. However, although the similarity to ancestral forms can vary from exact features to mere approximations, the novel pathways and forms used to accomplish these similarities are what make studies of morphological evolution worthwhile.

*Candidate gene approach*

The most studied cave-related phenotypes are eye regression and pigmentation [76-78, 80, 81]. Studies of the genetic bases of these changes were mostly approached from a developmental genetic viewpoint based on gene inhibition and over-expression methods, known as forward genetic analysis. The expression levels of several genes (*pax6, prox1, αA-crystallin, hsp90alpha, hsp90beta, pax2a, vax1, shh, twhh, ptc22, and nkx2.1*) known to be involved in eye developmental pathway has been compared between cave and surface forms [82-84]. In cavefish embryos only *pax6* showed lower expression in the optic cup, whereas *pax2a* and *vax1* had higher expression. A remarkable expanded expression of *shh, twhh, ptc22, and nkx2.1* was observed in the anterior mid-line of the fry [84]. An interesting example of a candidate gene is the αA-crystallin gene, whose coding and upstream regulatory sequences are identical in cavefish and surface fish [85, 86]. The anti-apoptotic factor *αA-crystallin* gene tends to be strongly down regulated in the lens of two different cavefish populations [86, 87]. In total, these studies

reveal that many candidate genes that are differently expressed between morphs, nevertheless it is still impossible to conclude which genes are the main causes of morphological change. Therefore additional genetic resources were necessary in the study of this organism.

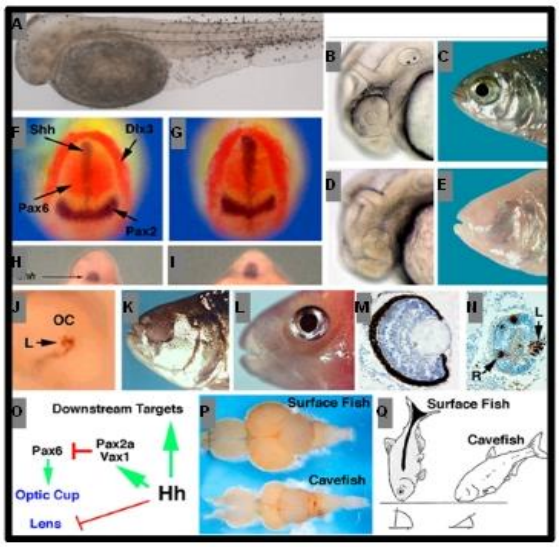



Figure 1. 3. Summary of phenotypes and expression studies in Astyanax Mexicanus. (A) Albinotic Pachón cavefish embryo with melanin positive melanoblasts after l-DOPA supplementation. (B, G) Surface fish embryo with eye primordium (F), and adult with eye (H). H, I) Pachón cavefish embryo with small eye primordium showing a reduced ventral sector (F), and eyeless adult (G). (H–K) Cavefish embryos show expanded *shhA* expression along the midline and contracted *pax6* expression in the eye fields. (H, I) Surface fish. (J, K) Pachón cavefish. (H, J). Neural plate stage viewed dorsally. (I, K) Ten somite stage viewed rostrally. Markers (*dlx3*) indicate the border of the neural plate and (*pax2*) boundary of the future midbrain and hindbrain region in the neural plate. (L, M) Overexpression of *shh* induces lens apoptosis (L) and a blind cavefish phenocopy (M) in surface fish. (N) Transplantation of a lens from an embryonic surface fish into the optic cup of a Los Sabinos cavefish rescues eye development. (O, P) Sections through embryonic surface fish and Pachón cavefish eye primordia showing apoptosis in the cavefish (P) but not the surface fish (O) lens. Arrows in (P) indicate the lens (L) and retina (R). (Q) Diagram showing effects of expanded Hh signaling in cavefish. (R) Comparison of brain morphology in surface fish and cavefish. Dorsal views with anterior on the left. (S) Differences in bottom feeding posture in surface fish and cavefish. (A) is reproduced from [81, 83, 84, 88-92].

*Quantitative trait analysis (QTL) studies in Astyanax mexicanus*

Most species studied for ecological or physiological traits lack genomic data and *A. mexicanus* is no exception. Even in the absence of a genome sequence this species became an exceptional model to study the genetic basis of morphological traits due to its amenability to genetic and QTL analyses [70, 76-78, 80]. The map was assembled using 259 markers to detect recombination frequencies and testing association of 12 traits that differ between cave and surface tetras (eye size (E), melanophore numbers (M), body condition rate (C), number of maxillary teeth (T), sensitivity to dissolved amino acids (A), rate of weight loss (W), body length (L), depth of the caudal peduncle (D), the placement of the dorsal fin (P), size of the SO3 dermal skull bone (S), numbers of anal fin rays (R), ribs number (B) [77]. Presence of numerous multi-traits QTL on the map such as EMT, EMCTW and EMRDS is very surprising given that these traits are not functionally related.

The QTL map of the other two crosses with Tinaja and Molino cave morphs with surface fish is also in the finishing state. According to preliminary data it is evident that the QTL maps while integrated do not map at the same positions as in Pachón, which supports the idea of independent evolution of the traits in the different caves [93]. Furthermore, crosses between geographically isolated cavefish populations (Molino, Pachón, and Tinaja) can produce progeny with a greater degree of eye development than that exhibited by either parent [67, 75].

The genetic map of *A. mexicanus* is a useful resource, which allowed for further candidate gene testing and the placement of some very important genes on the QTL map. Oculocutaneous albinism II (Oca2), the gene responsible for albinism in *Astyanax* cavefish, was the first candidate gene successfully mapped on the linkage map. Furthermore, this study illuminated that albinism arose by independent changes in Oca2 in three different cave populations [78]. In addition, for a number of cave related changes 12

candidate genes (Cg1, Fbp, Gh, Igf1, Igfbp5, Ins, Tfe3, Idh2, Oca2, Pax6, Shh, and Twhh were also placed on the map [77, 78, 81]. The phenotypic changes, even though much less studied also include constructive changes like increased complexity of feeding apparatus (larger jaws, mechano-sensory system (more taste buds, larger cranial neuromasts), and a more sensitive olfactory system and modified behavior patterns.

In addition to the conventional forward genetic approaches to better understand eye and pigmentation regression, some recent studies focused on behavioral changes associated with the cave environment. These experiments identified both sleep patterns and sensitivity to vibration as very important adaptations also observed in a multiple populations [94, 95].

*Impact of natural selection in the adaptation of cavefish phenotypes*

The impact of natural selection on cave evolved organisms and their regressive traits remains one of the most interesting questions. The neutral mutation hypothesis [96] was largely proposed for the eye regression in *Astyanax mexicanus*. It was suggested that given enough time and a sufficiently high mutation rate random mutations in eye-forming genes accumulate in cave animals under relaxed selective pressure [67, 97, 98]. Thus, eyes would eventually disappear because they are not necessary for survival in the dark environment. On the other hand adaptation theory attributes eyes regression to energy economy. Also, pleiotropic effects have been proposed in which sensory organs beneficial to survival in the cave environment are enhanced at the expense of eyes [77, 93, 99].

Only some experimental evidences have been obtained to give an idea about any of this hypothesis. QTL data provided information on the roles of natural selection *vs.* genetic drift in phenotypic evolution. The powers of those conclusions are obviously constrained by the power of QTL analysis [100]. For example, QTL analysis of the crosses between cave and surface individuals

showed that reduction of eye and of pigmentation occurred independently [77]. Therefore different evolutionary mechanisms could control regression of these traits. QTL assignment test that tests QTL polarity showed consistent polarity of all eye QTLs. This suggests that natural selection might be driving evolution of those loci. If both natural selection and drift are involved in the evolution of those traits, the QTL polarity should randomly change as it was observed for pigmentation QTL. These studies thus proposed natural selection driving eye loss as a main energy conservation mechanism. Contrary to that it was hypothesized that in pigmentation regression neutral mutation was the driving force [77]. However, the relationships between neutral mutation and selection are very poorly understood in *Astyanax* cavefish and further research on multiple natural populations is needed to clarify them.

### 1.3. Quantitative genetics approach in detecting genetic basis behind morphological traits

Another approach to detecting parallel or convergent genotypic adaptation in non-experimental systems involves quantitative genetic analysis (QTL). The major interest of QTL study lays in the identification of regions in the genome that harbor loci affecting complex trait variation and estimating the magnitudes and polarities of effects due to allelic variation at these loci in the experimental population.

QTL analysis is based on finding a trait of interest and performing crosses between individuals that differ phenotypically in that trait [101-106]. The next step is to develop and use multiple genetic markers that differ between the two phenotypically distinct populations in order to genotype the progeny. The patterns of segregation of these markers in the parents of a hybrid mapping progeny allow for the construction of linkage maps and the detection of QTL through correlations of genotype and phenotype in the mapping progeny [103, 106-108].

QTL is a powerful approach that allows us to ask many important questions in the evolution of morphological traits [8, 102, 104-106, 109]. For example, how many loci are responsible for the given trait and how much variance in phenotype may be explained by each locus? Are the loci of small or large effect? Are there interactions between the loci (epistatic effect) or are multiple traits encode by the same genes (pleiotropy)?

QTL studies are primarily used in model species because they are easiest to perform. For example, in *Drosophila melanogaster*, many genetic markers have been developed and the flies are easy to breed and maintain in the laboratory, which makes QTL analysis easier [110-112]. Nevertheless, QTL mapping had a lot of success in identifying genetic basis of morphological traits even in some non-model systems [6, 9, 78, 113, 114]. For example, stickleback marine and freshwater populations are radically different in their morphology; using QTL studies, multiple QTL were detected that controlled the numbers of gill rakers, lateral plates, pelvic spines, etc. [4, 6, 7, 9].

QTL mapping in non-model species is frequently combined with a search for candidate genes causing similar phenotypes in model species [78, 80, 115-117]. For example, a combination of genetic mapping and candidate gene determination was very successful in discovering the causes of albinism, alterations in the structure of melanin, and decreases in the numbers of melanophores in *Astyanax mexicanus* [78, 80]. Combination of QTL and candidate genes was also used in the QTL study to identify the genes and mutations responsible for morphological change in cave adapted isopod (*A. aquaticus*). Those genes fall into area of the genome responsible for presence *vs.* absence of pigmentation phenotype. However, the candidate genes were not causative once, but they are rather linked to the causative loci [118].

While QTL analysis provides a lot of information about the effects of QTL influencing the trait and their putative location, it gives little or no information about the molecular nature of the QTL [8, 119, 120]. The reason for

that is that fine scale QTL mapping is a prolonged and costly process of narrowing a QTL to a region with few enough candidate genes that each can be thoroughly tested. This ability to reduce QTL to a small number of testable candidate genes is necessary for increasing the rate at which QTLs are identified and proven. Thus, studies ideally need to combine both QTL and population genomics approach [5, 31, 32, 42, 44, 46, 119, 121].

Traditional linkage mapping is useful for identifying rare alleles and is not subject to the effects of population structure. However, loci that are identified by QTL mapping are specific to the parental lines of the experimental segregating populations and may not be representative of the genetic variation on which natural selection acts.

### 1.4. Inference of evolutionary history and demographic processes in the natural populations

Observed phenotypes in natural populations are mostly result of a delicate balance between selection and migration. Adaptive response appears to be modulated by gene flow and demographic history and can be predicted by divergence with gene flow models. Thus, the way towards understanding the genetic changes underpinning repeated phenotypic evolution require understanding the relative importance of selection versus historical processes [122-125]. Historical processes that generated patterns of phenotypic diversity in nature are particularly challenging to detect in natural populations [126]. Populations often exhibit heritable genetic differences that correlate with environmental variables, but the non-independence among geographically close populations (substructuring) complicates statistical inference of adaptation. Historical relationships among closely related populations could confound studies that compare different populations for adaptive processes [127]. Particularly for populations that have diverged recently, inferences regarding the role of environmental factors in driving evolution that assume

statistical independence among populations may be misplaced. Thus, when assessing adaptive phenotypic differentiation, the structure of evolutionary relationships among populations must be considered [128-130].

To understand maintenance of variation with species and importance of local selection and demography, determination of population parameters such as genetic distance, gene flow, population structure and effective population size is very important [131-133]. In particular, it is important to examine whether patterns of adaptive morphology observed within populations are replicated across the natural range of the species. This information will allow us to ask the following question: Are those phenotypes derived from a single event followed by dispersal, or if they are adapted multiple independent times? Is there a gene flow between the independently derived lineages? In other words, do adaptive phenotypes have a single evolutionary history (parallelism) or they appeared by convergence (from the distantly related lineages)? Having this information it can be determined when and under what circumstances phenotypic variations has evolved and separate recent selective events from historical processes.

*Population structure*

Despite the success of population genetics in modeling neutral variation, many assumptions of these predictions are frequently violated in natural populations, which make the estimation of genetic parameters a challenging task. One of the very important aspects of describing demographic effects is to determine population substructure and gene flow. This information is necessary for accurate estimates of effective population sizes ($N_e$), genetic diversity, and migration rates, which are key parameters in describing the dynamics and relationships among different populations in the wild. Many natural populations consist of partially isolated local subpopulations that vary in size and structure; with varying patterns of gene flow among them, individuals from the same

geographic region may mate with each other; and as such, they cannot be accounted for easily [130, 134] [123, 133]. Clearly, it is impossible to model the full biological complexity of demographic events, so we must look for the simplest models that capture the relevant features. Typically, we ask the question: How can we detect deviations from the null model and how can we estimate some of the important quantities related to the demographic models?

Population structure separates individuals into distinct reproductive units. Each of these units may behave like an ideal population Wright-Fisher population; finite size, where each individual contributes an infinite number of gametes to a gamete pool, and then each member of the next (finite) generation is drawn from that gamete pool) [135]. Over time, the stochastic nature of the evolutionary process will lead to genetic differentiation between populations; different allele frequencies among the populations or even complete fixation of different alleles in different populations. For example, if there was a locus with two alleles in multiple sub-populations, each with the frequency pi; the expected average heterozygosity ($H_e$) in the combined population under Hardy-Weinberg equilibrium (HWE) would be 2p (1-p), where p represents average allele frequency [136]. However, because of the reproductive isolation between these sub-populations, the heterozygosity will be reduced by the amount of allele frequency variance across sub-populations. This variance in allele-frequency is directly related to the population inbreeding; the frequency of heterozygotes compared with that expected when genotypes are in HWE. Also, inbreeding increases relatedness between the individuals based on common ancestry; “identity by descent” (ibd) [137]. Thus population structure can be defined as ibd and it can be measured as relatedness between the individuals relative to the populations and between populations relative to the species. These measures were proposed by Wright and are termed F-statistics with the different hierarchical levels denoted as $F_{IS}$ (the mean reduction in heterozygosity of an individual due to non-random

mating within a subpopulation; i.e. genetic inbreeding within subpopulations), $F_{ST}$ (the mean reduction in heterozygosity of a subpopulation relative to the total population due to genetic drift among subpopulations; i.e. between-population differences) and $F_{IT}$ (mean reduction in heterozygosity of an individual relative to the total population) [128, 137-139].

Wright's F statistics measure the correlation between alleles drawn at different levels of a subdivided population [139, 140]. Evolutionary processes, such as mutation, migration, inbreeding, and natural selection influence that correlation. The original definition of F-statistics was to measure the amount of allelic fixation owing to genetic drift.

Wright's $F_{ST}$ model is an idealized n-island model in which an infinite number of populations receive immigrants from an infinitely large mainland population [141]. $F_{ST}$ is a simple function of effective population size and the migration rate $F_{ST} = 1 / (1+4N_em)$ in which the strength of genetic drift is proportional to 1/N, while the strength of gene flow is proportional to m. When $F_{ST} = 0$ there is no differentiation between the populations, when $F_{ST} = 1$ differentiation is maximum.

Levels of population differentiation are typically quantified using a variant of Wright's $F_{ST}$ parameter, which measures the proportion of variation in a sample that is distributed among populations. This has been the most used approach, partly because of its robustness and partly because it is simple to implement [139]. This estimator of $F_{ST}$ can give us a good idea about what form of demographic model may apply to the data.

However, the relationship between real demographic parameters and $F_{ST}$ are not so simple. $F_{ST}$ is a very good estimate of the population differentiation; however the estimates of $F_{ST}$ should not be directly translated into $N_e$ x m; measure of gene flow. The reason for that is that the relationship of the variance in gene frequencies among different populations ($F_{ST}$) is related to the number of migrants is non-linear function of $N_e$ x m [123, 142]. This non-

equilibrium demography typically increases the variance of summary statistics, highlighting the importance of using simulations to study power and efficiency of this approach [143].

Population genetic theory [144-146] has allowed for the development of more robust methods to measure population structure including coalescent-based likelihood methods (where likelihoods are estimated by stochastic simulation) and Bayesian methods [128, 147, 148]. These methods account for more of the underlying biology of populations. For example, they give insights into the rates of mutation and migration; because of that they provide more information on population structure than $F_{ST}$ summary statistics [128, 129].

Estimates of $F_{ST}$ using the method of moments and Bayesian methods have not been extensively compared. An example of the differences among calculating method of moments, maximum-likelihood and Bayesian estimates of F-statistics has been shown by Holsinger & Weir [128]. They used a study on human populations in which the allele frequency differences at blood group loci were measured (Table 1.2). Based on this example it has been suggested that those differences in estimates are small under the following conditions: when the average number of individuals per population is moderate to large (>20), when the number of populations is moderate to large (>10–15) and when most populations are polymorphic.

| Parameter | Method of moments | Maximum likelihood | Bayesian |
|---|---|---|---|
| f | 0.03090 | 0.03460 | 0.05030 |
| theta | 0.00402 | 0.00640 | 0.01890 |
| F | 0.03480 | 0.04080 | 0.06830 |

f-Coancestry for alleles within an individual relative to the subpopulation in which it occurs; equivalent to Fis; theta-Coancestry for randomly chosen alleles within the same subpopulation relative to the entire population; equivalent to $F_{ST}$; F-Coancestry for alleles within an individual relative to the entire population; equivalent to Fit.

Table 1.2. Differences among calculating method of moments, maximum-likelihood and Bayesian estimates of F-statistics. Data from a classic study on human populations that investigated the allele frequency differences at blood group loci; from [128].

However, when the assumption of uniform effective population sizes ($N_e$) and symmetric migration rates assumed by summary statistics is violated, there is big discrepancy between the above methods and parameter estimates. This is especially evident in populations with high $N_e$ [149, 150] that are weakly structured, and when highly polymorphic molecular markers are used for the structure detection [151]. An example of that is reported in the estimation of the population structure of big eye tuna populations. In this study Bayesian cluster analyses, and coalescence-based migration rate inferences supported high migration rate and lack of genetic structure, which contrasts the $F_{ST}$ estimates [152].

Population history and demography also has an important role in the mode (balancing and directional selection; see description latter) and efficacy of natural selection. For example, low levels of gene flow enhance local adaptation [153]. On the other hand, efficacies of positive and negative selection are reduced in populations with small effective sizes or those that experienced severe bottlenecks [123]. Given those effects, the details of the demographic processes are not only important for the null model against which selection will be tested, but they are also very important for the appropriate models of selection.

### 1.5. Selection detection in the natural populations

Detecting selection in the wild has been a very challenging task for a long time [131, 154-157]. The main problem lays in distinguishing the effects of selection on particular loci from the background demographic history of the population including changes in population size, migration and divergence. A range of approaches has been used to identify regions that are likely to have been targeted by selection. The principle behind most of the approaches is that only genome-wide effects can inform us about demography and phylogenetic history of the populations that is due to neutral loci. Contrary to that, locus-

specific effects help identify genes that are important for fitness and adaptation and will often reveal different patterns of variation (reviewed in [157]). Selected loci could be detected based on reduced diversity, excess of linkage disequilibrium (discussed later) within and among populations, haplotype structure, fixed DNA sequence divergence among related taxa or high geographic differentiation relative to the other loci across the genome.

Besides detecting selected loci it is also very important to distinguish between the modes of the selection present in the surveyed region. The modes of selection that are commonly identified in the natural populations either operate on the mean value of the trait (balancing selection) or on the one extreme of the trait distribution (positive selection). The actions of selective forces are than reflected in the allele-frequency of the adapted populations [155]. Balancing selection causes multiple alleles to be maintained in a population, often at fairly constant frequencies [158]. In the case of positive selection, favorable mutation becomes fixed in a population.

Different tests for selection are designed to find different departures from “neutral” (demographic) model since all demographic complexities cannot be taken into account, as discussed before. In the other words, there is no such a thing as “perfect test” for selection. How can we make sure than that we have identified selected locus? Repeated detection of the same loci by different statistical methods could be one of the solutions. Another possibility is to investigate biological replicates that inhabited different environments and experienced the change in selective pressures, which will have a strong power to understand genome-wide adaptation. Here, we are presenting approaches for selection detection that are commonly used in non-model systems [155, 159].

*Outlier detection*

Outlier analyses methods became definitely beneficial as a preliminary method

with the ability to screen numerous markers in genome scans to identify candidate genes for further investigation. In general, these methods consist of identifying loci that differ from expectations under the neutrality based on summary statistics ($F_{ST}$ and homozygosity). The power to distinguish outlier loci from neutral loci is dependent on the null distribution of the summary statistics across the genome. The null distribution can be experimentally obtained by collecting hundreds of loci. However, in non-model system it is rare to have so many loci, thus simulations to model neutral loci are frequently used [154, 160-162].

The problem is therefore to generate the distribution of statistics under demographic model congruent with the observed data. The models that are used in outlier tests can involve different population structures and histories and can assess the influence of different demographic and non-equilibrium scenarios. These models are robust to a wide range of alternative demographic models. It is likely that they will detect outliers with unusually high or low $F_{ST}$ and will identify selection at one or many loci through pairwise comparisons of populations [154, 156, 157, 163-166]. The basic rationale for testing natural selection using these methods is that loci influenced by positive selection will show a larger genetic differentiation than neutral loci (high $F_{ST}$). On the other hand, loci that have been subject to balancing selection will show a lower genetic differentiation (low $F_{ST}$). However, outlier tests typically generate discrepancies when numbers of immigrants per generation are unequal, the true population history consists of repeated branching events, or the connectivity of populations is uneven [167-169]. These discrepancies are mostly reflected in limited power in detecting balancing selection. Isolation, population bottlenecks, and heterogeneity of populations are also increasing the possibilities to detect false positive or negative loci [154] as well as weak divergent selection [170].

A recent study compared different outlier methods using simulated data

where the selected loci were defined. They showed different sensitivity to detect balancing selection. Again, accurate detection of balancing selection is an inherent weakness of outlier approaches. A similar conclusion was reached in the recent study of adaptation in natural populations of sticklebacks [169]. This study used multiple outlier tests in order to compare selected loci across the methods. They were able to consistently indentify the same loci across the multiple methods. However, loci under balancing selection showed major issue of the methods. Both of these studies point out that statistical methods should be carefully chosen based on the purpose of the study with special attention to the error rates of the different methods [168].

*Linkage disequilibrium and haplotype based tests*

Using information from only a single marker to make inferences about natural selection clearly ignores an important source of information; namely non-random associations of alleles at linked loci (i.e. linkage disequilibrium (LD). So, instead of asking whether there is a specific pattern of genetic variation on the single SNP, one can extend that on the many correlated markers [133, 155, 171-178]. However, this is dependent on the availability of the data. The main problem in the studies of non-model organisms is that one will rarely have information about the order of the markers or distance between them. In the species in which this information is available we can gather the information of multiple markers in the same region and contrast different region in the genome in order to find selected locus. The association between alleles across loci is defined as linkage disequilibrium (LD) and it has been traditionally calculated as a function of a pair of loci [176, 179]. LD between alleles is defined as DAB = pAB – pA x pB which represents difference between the frequency of gametes carrying the pair of alleles A and B at two loci (pAB) and the product of the frequencies of those alleles (pA and pB) (reviewed in [173, 180]).

These classical definitions of LD are very important and widely used, but its patterns are well known for being noisy and unpredictable. For example, two pairs of markers can be in complete disequilibrium even when they are unlinked whereas LD for the pair of markers next to each other might be weak (reviewed in [176]). Also, these statistics consider only two loci at a time, whereas we may be interested to calculate the extent of LD across a chromosome segment that contains multiple markers. LD in genotypic data can be quantified [178, 180], but the lack of information about the haplotype phase weakens the signal of nonrandom association sufficiently that this approach is not often taken. Haplotypes are not known in unrelated individuals and they have to be inferred, which is easily done based on frequencies from the genotypes in the surveyed populations [181]. They are typically more informative than individual genotypes, when LD between the phased markers is strong. The strength of haplotypes is in the usage of multiple SNPs together (haplotype blocks) while estimating population genetic parameters [177, 182, 183]. Thus, it is more common to use a statistical method based on population genetics theory to infer haplotype phase from genotypic data and then to treat the inferred haplotypes as if they were data.

Haplotype diversity can be the result of migration, mutation, selection, small finite population size (eg. [108] and those effects can be inferred using many different methods [155, 158, 182-184]. For example, we can measure haplotype diversity using a count of the number of observed haplotypes in a region or by the expected haplotype heterozygosity based on haplotype frequencies in a region [176, 177]. One of the simplest approaches when data for more than one population are available is to partition the haplotypes into contributions within and among populations. This partitioning first suggested by Ohta [174, 175] and it is similar to Wright's statistics that partitions variability within the populations ($F_{IS}$) and between the populations ($F_{ST}$) (*for explanation see demographic section*) [185]. When comparing data that way one can find

the specific haplotypes in the population that diverged from the other haplotypes, just like the single outlier SNP (reviewed in [176]). If the natural selection favored adaptation to local conditions we will detect increased $F_{ST}$ whenever alleles at different loci are favored [156, 157]. Partitioning haplotypes in different regions in the genome is an appropriate step when trying to determine whether differences in haplotype frequencies result from natural selection stemming from differences in conditions among populations.

*Integrative approach in detecting adaptive loci behind repeated phenotypes*

In the past few decades LD has been also utilized as a tool for genetic mapping of trait or disease loci in the natural populations [172, 173, 176, 186]. Mapping based on LD requires alleles to be in LD with an allele responsible for a quantitative trait, across the entire population. To be a property of the whole population, the association must have persisted for a considerable number of generations, so the marker(s) and causative locus must therefore be closely linked. LD mapping and its variations (e.g. association mapping, selective sweep mapping) are commonly used approaches in finding genes that underlie ecologically important traits in natural populations [143, 172, 176, 177, 183]. These approaches rely on a statistical association between genotype and phenotype, and have shown great potential for fine mapping of traits and for identifying functional markers [171, 173, 186-188].

LD mapping implies that there are small segments of chromosome in the population that will carry identical haplotypes if there is a QTL somewhere within the chromosome segment. Therefore if individuals from adapted population carry the same haplotypes, which are likely at a point of the chromosome carrying a QTL, then their phenotypes and genotypes are correlated. This approach that was originally developed in *Zea mays* combines the advantages of traditional QTL mapping (low marker density requirements and high allele richness) and LD mapping (high mapping resolution and high

statistical power) [189].

Unfortunately, due to the unknown positions of the markers in the genome these studies are not very common in natural systems. A successful example of similar approach (only individual SNPs were used instead of haplotypes) that combined individual markers from natural populations survey with QTL from experimental crosses was given by Rogers and Bernatchez [32]. They combined population scans of the genome for “outlier” loci with QTL mapping to examine the genetic basis of growth rate differences between limnetic and benthic ecotypes of whitefish. By performing QTL mapping using AFLP markers that were previously used in population genomics scans [31, 190] they were able to determine that the loci closest to growth rate QTL were the same as loci showing elevated differentiation in genome-wide scans of natural populations. Combinations of genome-wide scans and traditional QTL mapping allows for testing whether QTLs identified in different populations have played a part in adaptive phenotypic differentiation. This is clearly a powerful tool in non-model organisms in which physical genome is not available [8, 31, 42, 44, 46, 106, 155].

### 1.6. Objectives

Identification of the causative polymorphisms underlying parallel and convergent phenotypic traits and understating the evolutionary forces driving that change is an extremely challenging task in organisms, which lack extensive sequence information and genomic resources. The main goal of this work was to contribute to the general understanding of the evolutionary mechanisms underlying phenotypic evolution in natural populations using a non-model organism from the following perspectives:

1. To determine the history of cavefish populations and the independent evolution of phenotypic change by examining population structure, the rates of gene flow among populations and effective population size using multiple

independent microsatellite markers across different populations.

2. To develop the new genomic tools to disentangle the contribution of genetic drift, natural selection, new mutations and pre-existing variation in evolution of morphological traits in *Astyanax mexicanus*.

# CHAPTER 2

## Gene flow and population structure in the Mexican blind cavefish complex (*Astyanax mexicanus*)

RUNNING TITLE: Demography of cavefish and surface conspecifics

Submitted manuscript

***Adapted from**

Martina Bradic[1, 2], Peter Beerli[3], Francisco J. García-de León[4], Sarai Esquivel-Bobadilla[4], Richard L. Borowsky[1]

[1]Cave Biology Group, Biology Department, New York University, NYC, NY, USA

[2]Univesidade Nova de Lisboa, Oeiras, Portugal

[3]Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

[4]Centro de Investigaciones Biologicas del Noroeste (Laboratorio de Genética para la Conservación), La Paz, Baja California, Mexico

## 2.1. SUMMARY

The evolutionary forces driving convergent evolution in natural populations remains poorly understood. To better understand these mechanisms, we studied multiple populations of Mexican Blind Cavefish (*Astyanax mexicanus)*, exhibiting very similar cave related phenotypes, including decreased eye size and pigmentation and compensatory hypertrophies of other senses. Here, we ask how many times those phenotypic traits have evolved independently across the three distinct geographical areas of their range. We assessed genetic structure and differentiation within and among the populations. The widespread surface localities are, with some exceptions, genetically similar to one another, whereas the cave populations are differentiated and have at least five distinct origins in the three main regions. We find lower genetic diversity in cave populations than in related surface populations due to the smaller effective population sizes probably because of the food and space limitations. However some of the cave populations receive migrants from the surface and exchange migrants with one another, especially when geographically close. This admixture results in significant heterozygote deficiencies at numerous loci due to Wahlund effect. In cave populations receiving migrants from the surface, we identified small numbers of individuals that are both phenotypically and genotypically intermediate, affirming gene flow from the surface. This study provides the evidence that the phenotypic changes have evolved independently multiple times as well as an important role of natural selection in driving the phenotypic divergence despite the gene flow.

## 2.2. BACKGROUND

The mechanisms underlying the evolution of similar phenotypes in independent natural populations pose a long-standing question in evolutionary biology. Apart from a few examples [191, 192] the molecular nature of convergent phenotypes remains largely unknown. Also unknown is the extent to which new mutations versus preexisting genetic variation in ancestral populations contribute to convergence (or parallelism) [4, 193].

Convergence is of interest to evolutionists for several reasons, one of the most important of which is that it provides an element of replication to evolutionary studies that is often otherwise absent. Replication allows for the powerful testing of evolutionary hypotheses. Cave-dwelling organisms provide the best known examples of convergences, sharing similar phenotypes such as loss of eyes and pigmentation across diverse taxonomic groups [194].

The Mexican blind cavefish (*Astyanax mexicanus*) is nearly unique among cave animals because the cave forms have closely related surface conspecifics and the two forms are fully interfertile [67]. The ability to hybridize the cave and surface forms permits the genetic analysis of the factors involved in cave adaptation. There are 29 known cave populations of this species dispersed over a broad geographic range and the group may present multiple examples of convergence.

Each population inhabits a food and light restricted cave environment; their members exhibit numerous cave-related evolutionary changes, including reduction in pigment and eye size, hypertrophy of non-optic sensory organs, increased condition factor, and robust patterns of reduced sleep; presumably all are evolved in response to reduced food availability in caves [65, 67, 76-78, 94].

Thus, the cave colonizations of *Astyanax* populations provide replicates of an excellent “natural experiment” which allows us to address important

evolutionary questions, including the extent to which morphological, behavioral and physiological evolution is driven by selection versus drift [79, 81, 97]. These two alternatives can be distinguished in a number of ways in this system, but any determination will require an understanding of the underlying demography of the populations as well as a clarification of the relationships among them.

Previous phylogeographic studies of Astyanax cavefish, using microsatellite and mtDNA, showed that the cave populations are derived from at least two different surface stocks that inhabited the Sierra de El Abra and nearby regions in succession [71-74]. The estimates from mtDNA suggest that these two groups diverged about 6.7 Mya [72]. Surface forms of the older stock originally inhabited the rivers in the El Abra region and were the likely ancestors of a series of cave populations, which we designate as “old.” Subsequently, the surface fish of the old stock went locally extinct. The region was then invaded by another stock of A. mexicanus; its descendants are the current occupants of the region’s surface waters and a second set of cave populations we designate as “new.”

While previous studies revealed that the extant cave populations were derived from a minimum of two ancestral stocks, there may have been more. In addition, the question of how many independent invasions of the underground led to the present day *Astyanax* cave fauna remains unanswered. To understand the demographic component of the phenotypic evolution we studied cave populations from the full extent of their known distribution. Our study includes 11 populations of cavefish, some not previously studied, as well as 10 populations of surface fish from the surrounding area. We give a detailed description of genetic differentiation in multiple cave populations and their relatedness with surface morphs, and estimate effective population sizes and the rates of gene flow among select populations based on multiple independent markers.

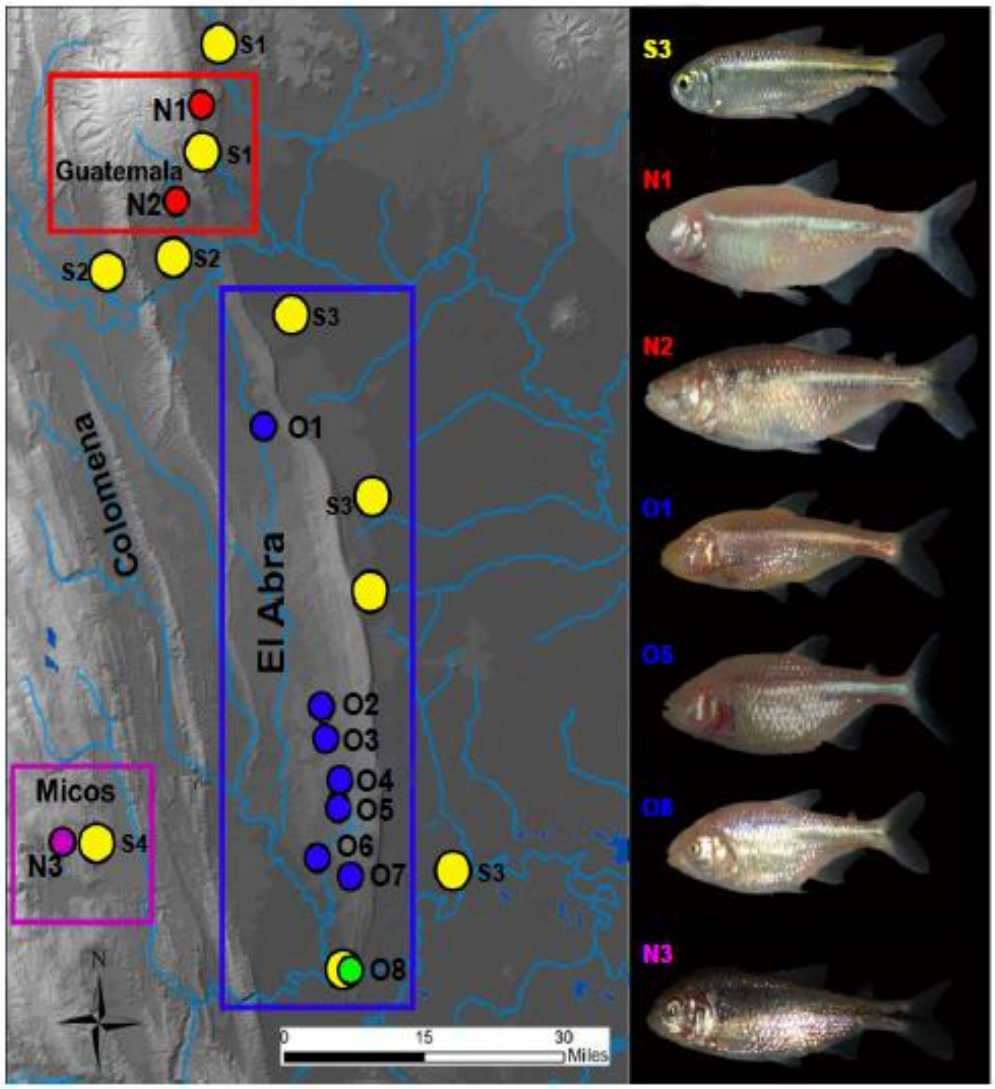



Figure 2.1. Map of the Sierra de El Abra region showing all the cave and surface collection sites. Colored lines delineate major geographical regions (labeled below), as follow: El Abra region: O1 – O8 (blue & green circles); Guatemala region: N1 – N2 (red circles); Micos region: N3 (purple circles); Surface localities S1 – S4 (yellow circles).

## 2.3. RESULTS

### Genetic Diversity

We calculated descriptive statistics using 26 unlinked microsatellite markers. The number of alleles and proportions of polymorphic loci were generally higher in surface than in cave populations, although there was considerable variability among populations (Table 2.1). Genetic variability was significantly lower in the cave populations than in the surface (Table 2.1). Average allelic number (Ar) ranged from 2.25 ± 0.50 in the Guatemala region (N1, N2) to 2.54 ± 0.26 in the El Abra (O1 to O8), to 3.63 ± 0.14 in the surface populations. Surface Ar was significantly greater than Ar in the Guatemala and in the El Abra ($t_{16}$ = 11.6, $t_{10}$ = 8.7, respectively, $P < 10^{-6}$ for both). Micos cave (N3) which is known to have both cave and surface-dwelling phenotypes from the previous studies [65, 73] had an intermediate average number of alleles per locus of 2.98. We also detected monomorphic loci (NYU26, 26C, 218A, 213B), which shared the same alleles among El Abra cave populations (data not shown here). Unbiased *e*xpected heterozygosities ($H_e$) were also higher and significantly different (0.82 ± 0.04) in surface populations than in the El Abra (0.55 ± 0.07; $t_{16}$ = 10.8, $P < 10^{-6}$) or Guatemala (0.49 ± 0.13; $t_{10}$ = 8.7, $P < 10^{-5}$) populations, while the Micos population (N3) exhibited intermediate heterozygosities of 0.66 (Table 2.1).

| Region | Population | Code | N | n | P | A | Ar | He | Ho | Lat | Long |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **EL ABRA** | Pachón | **O1** | 45 | 36.31 | 0.92 | 4.81 | 2.19 | 0.48 | 0.47 | 22.60 | 99.05 |
| | Yerbaniz | **O2** | 12 | 9.46 | 0.96 | 6.00 | 2.67 | 0.60 | 0.60 | 22.20 | 98.97 |
| | Japonés | **O3** | 10 | 7.77 | 0.92 | 4.12 | 2.55 | 0.57 | 0.55 | 22.10 | 98.95 |
| | Arroyo | **O4** | 12 | 9.19 | 0.88 | 3.62 | 2.36 | 0.53 | 0.50 | 22.20 | 98.97 |
| | Tinaja | **O5** | 4 | 3.56 | 0.88 | 2.72 | 2.5 | 0.56 | 0.58 | 22.08 | 98.95 |
| | Curva | **O6** | 13 | 10.00 | 0.88 | 3.75 | 2.3 | 0.43 | 0.49 | 21.98 | 98.93 |
| | Toro | **O7** | 3 | 2.69 | 0.77 | 2.23 | 2.98 | 0.60 | 0.55 | 21.85 | 98.93 |
| | Chica | **O8** | 119 | 104.08 | 1.00 | 9.27 | 2.74 | 0.64 | 0.60 | 21.85 | 98.93 |
| | Mean statistics of the group | | 27 | 22.88 | 0.90 | 4.56 | 2.54 | 0.55 | 0.54 | - | - |
| **GUATEMALA** | Molino | **N1** | 22 | 19.31 | 0.85 | 3.04 | 1.89 | 0.40 | 0.39 | 23.06 | 99.16 |
| | Caballo Moro | **N2** | 26 | 22.69 | 1.00 | 5.73 | 2.6 | 0.58 | 0.53 | 22.92 | 99.20 |
| | Mean statistics of the group | | 24 | 21 | 0.92 | 4.38 | 2.25 | 0.49 | 0.46 | - | - |
| **MICOS** | Subterráneo | **N3** | 72 | 58.88 | 1.00 | 10.96 | 2.98 | 0.66 | 0.57 | 22.10 | 99.18 |
| **SURFACE** | Río Frío | **S1** | 10 | 7.08 | 1.00 | 6.64 | 3.71 | 0.72 | 0.73 | 22.99 | 99.15 |
| **STREAMS** | Arroyo Sarco | **S1** | 32 | 27.42 | 1.00 | 11.62 | 3.57 | 0.82 | 0.82 | 22.02 | 99.32 |
| | Chamal | **S2** | 13 | 7.38 | 0.96 | 6.62 | 3.35 | 0.83 | 0.75 | 22.84 | 99.20 |
| | Río Meco | **S2** | 27 | 19.27 | 1.00 | 10.00 | 3.67 | 0.81 | 0.74 | 22.82 | 99.31 |
| | Río Tantáon | **S3** | 28 | 21.50 | 1.00 | 11.96 | 3.52 | 0.85 | 0.74 | 22.37 | 98.90 |
| | Río Florído | **S3** | 15 | 9.60 | 1.00 | 7.92 | 3.53 | 0.80 | 0.73 | 21.98 | 98.77 |
| | RióTampaón | **S3** | 26 | 17.23 | 1.00 | 10.15 | 3.74 | 0.84 | 0.77 | 21.85 | 98.94 |
| | Río Santa Clara | **S3** | 24 | 19.62 | 1.00 | 11.04 | 3.7 | 0.85 | 0.82 | 22.50 | 98.9 |
| | San Rafael Los Castros | **S3** | 25 | 19.62 | 1.00 | 11.19 | 3.65 | 0.84 | 0.76 | 22.75 | 99.02 |
| | Rio Subterráneo Valley | **S4** | 30 | 22.88 | 1.00 | 11.08 | 3.83 | 0.83 | 0.76 | 22.13 | 99.17 |
| | Mean statistics of the group | | 23 | 17.16 | 1.00 | 9.82 | 3.63 | 0.82 | 0.76 | - | - |

Table 2.1. Sample information and summary statistics of the sampled populations. N = sample size per population; n = mean sample size over all loci; P = proportion of polymorphic loci; A = mean number of alleles per locus; Ar = mean number of alleles standardized to the smallest sample. $H_e$ = unbiased expected heterozygosity standardized according to the (2N/2N-1))*$H_e$) formula; $H_o$ = observed heterozygosity, Lat is the latitude and Long is the longitude.

**Genotypic frequencies**

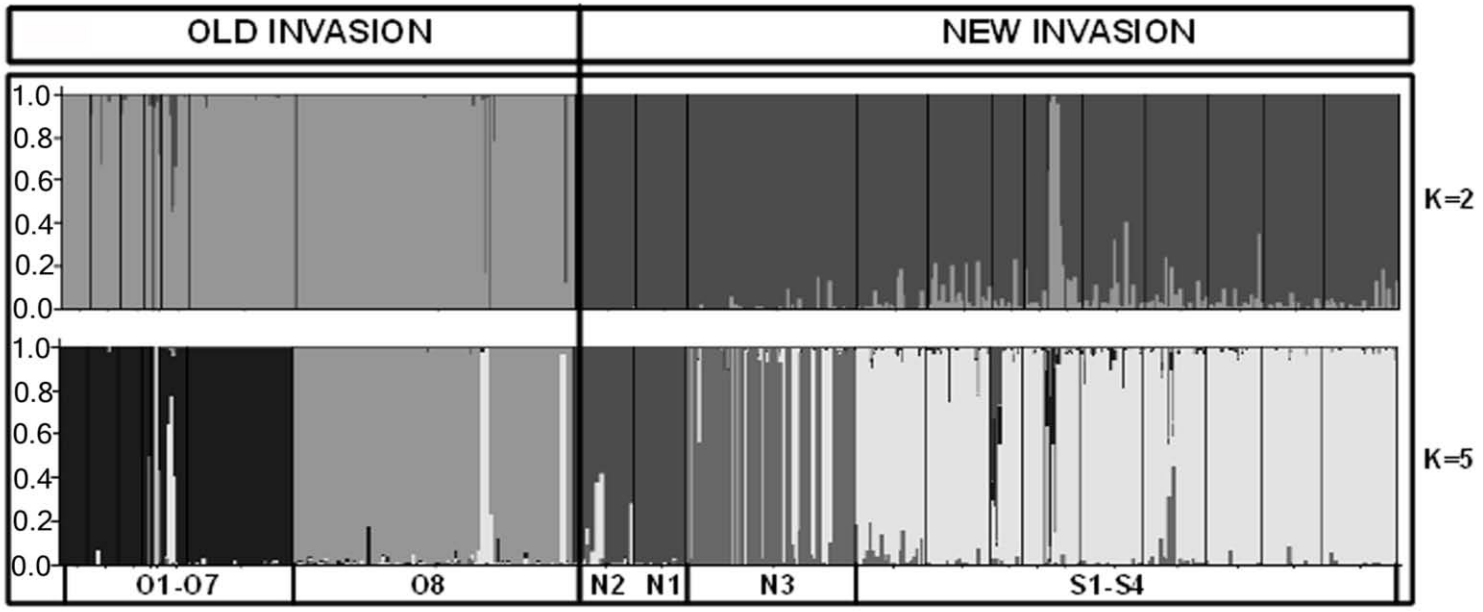
We performed 519 tests (27 values were excluded due to missing or monomorphic data) and detected 71 significant departures from HWE (based on 0.05 level of significance and standard Bonferroni corrections). Most of the significant loci showed heterozygote deficiency characterized by a positive $F_{IS}$ value. Heterozygote excess was detected in a few, mostly surface, populations for five loci (214D, 210A, 202D, 104A and 241B). Significant deviations from HWE were mostly present in the populations that were previously described as phenotypically mixed [64, 65]: El Abra (O8) and Micos (N3) populations (13 and 16 loci out of 25 scored, respectively. Presumably, this reflects population subdivision. The other cave populations exhibited only small numbers of loci out of HWE and these differed from one population to the next. One locus (213B) was out of HWE in many populations (9 out of 21), which may reflect the presence of null alleles at this locus.

**Population structure analysis and differentiation**

As a starting point to infer the relationships among populations we used the clustering algorithm implemented in the program STRUCTURE [130]. We explored different numbers of populations K to uncover hierarchical population structure. The clear distinction among the two groups when K = 2 is consistent with the hypothesis that all of the populations we studied originated from two stocks: a “new” stock including present-day surface-forms and the “new” cave populations from Micos (N3) and Guatemala region (N1,N2), and an “old” stock including the El Abra cave populations (O1-O8) and their extinct progenitors. Further structuring represents divergence of O8 from the other El Abra populations at K = 3, the more recent divergence between the surface populations (S1 - S4) and the new cave populations (N1 – N3) at K = 4, and the separate origins of the new cave populations at K = 5. Optimal K (Evanno et al. (2005) estimated the most likely number of populations at K = 5 (Figure 2.2B). We performed a STRUCTURAMA analysis, which estimated the same

value of K = 5 (posterior probability of 90%; results not shown).

These five independent groups of the populations are: 1) El Abra caves (O1-O7), 2) El Abra cave mixed population (O8), 3) the new cave populations to the north in the Guatemala region (N1, N2), 4) the new cave mixed population in southwest Micos region (N3), and 5) the surface populations (Figure 2.2A and 2.2B). The STRUCTURE analysis also revealed that four of the cave populations (N3, O8, N2 and O2) contained alleles from surface populations at several loci while the surface populations showed a smaller number of the alleles from the caves.


OLD INVASION
NEW INVASION
1.0
0.8
0.6
0.4
0.2
0.0
K=2
K=5
O1-O7
O8
N2
N1
N3
S1-S4

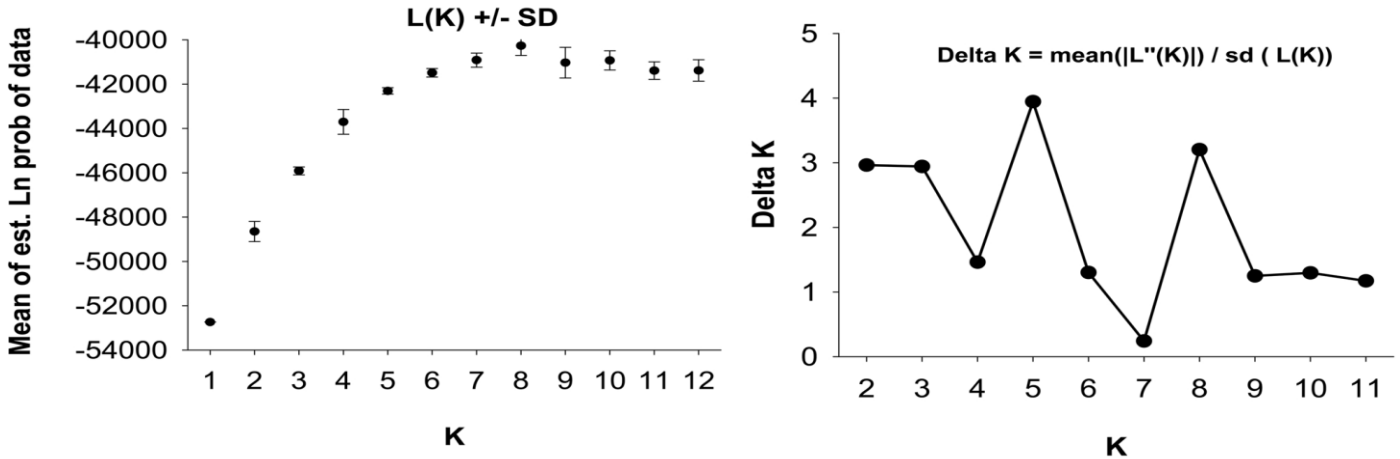
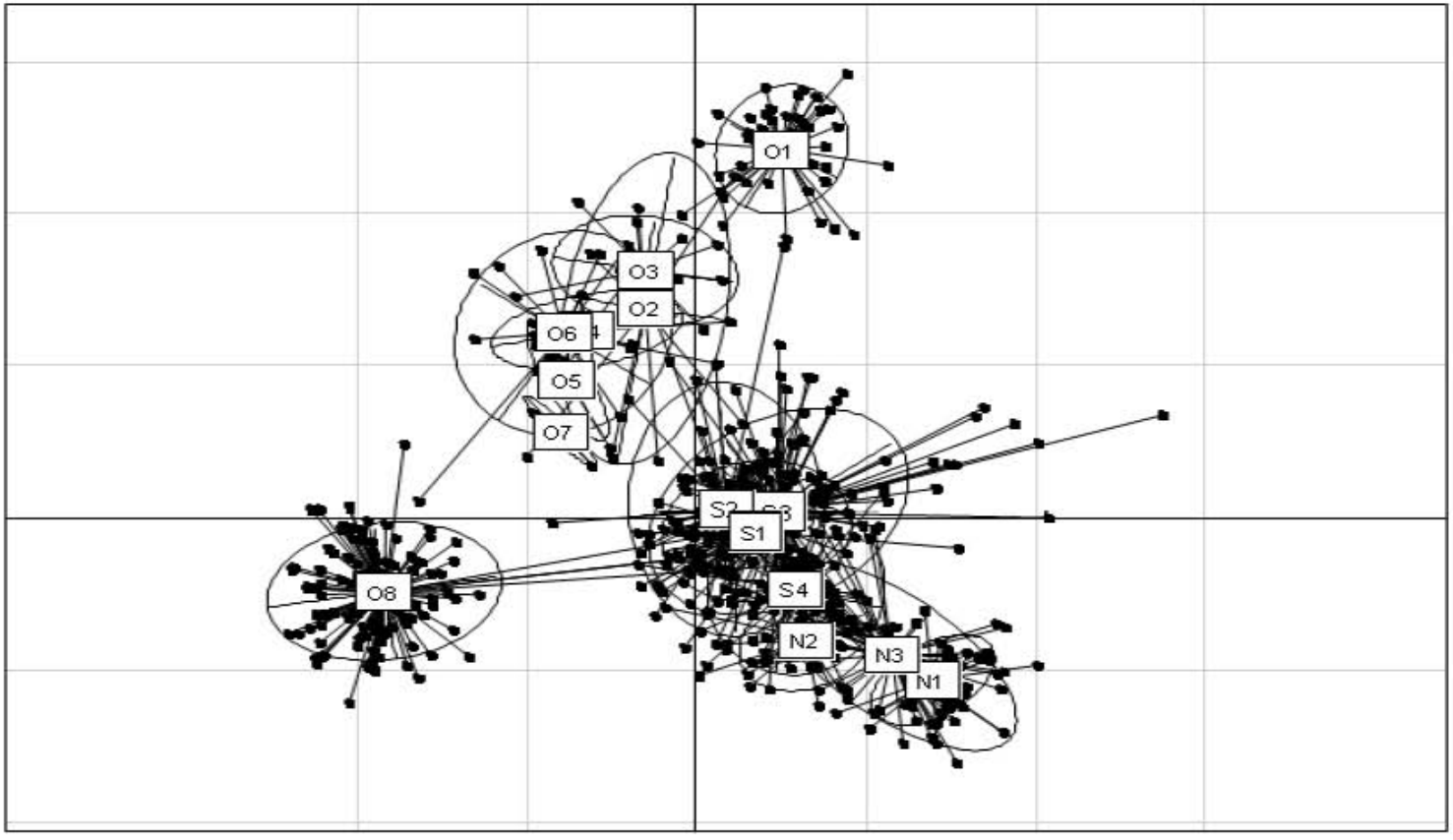



Figure 2.2. Estimated population structure of *Astyanax* cave and surface populations using STRUCTURE for K = 2 and K = 5 population groups. A. Each individual is represented by a thin vertical line, which is partitioned into K segments that represent its estimated population group membership fractions. Black lines separate individuals from geographical site locations (labeled below), which are as following: El Abra: O1 – O7; Chica (O8); Guatemala: N1 – N2; Micos: N3; Surface: S1 - S4. B. Mean posterior probabilities of ten runs for each K, K = 1 to K = 12. C. *K* = 5 had the highest Δ*K vs. K* peak height [195].

We further tested the genetic distances among populations using the metric of shared alleles. Figure 2.3A illustrates that the entire El Abra cluster is the furthest away from the cluster of the “new” caves (N1, N2 and N3). Genetically, one El Alabra cave population (O8) was equidistant from the “old” and “new” lineages, while the Micos (N3) cave population shared the most alleles with surface populations (Figure 2.3A).

Private allele estimates were calculated based on groupings of populations united by geographical proximity, which also corresponded well to the groupings revealed by STRUCTURE and shared allele distances. In addition, the private allele content is significantly higher in surface compared to cave populations (Figure 2.3B) (minimum $t_{49} = 4.23$, $P < 0.0001$). The shared alleles and private allele proportions between surface and cave populations (Figure 2.3A and B) suggests that the allelic contents of cave populations are largely subsets of alleles of the surface stock. Thus the observed variation in the caves is mostly the result of standing genetic variation from the ancestral surface stock as well as possible gene flow between the populations (Figure 2.3B).


O1
O3
O2
O6
O5
O7
S1
S4
O8
N2
N3
N1

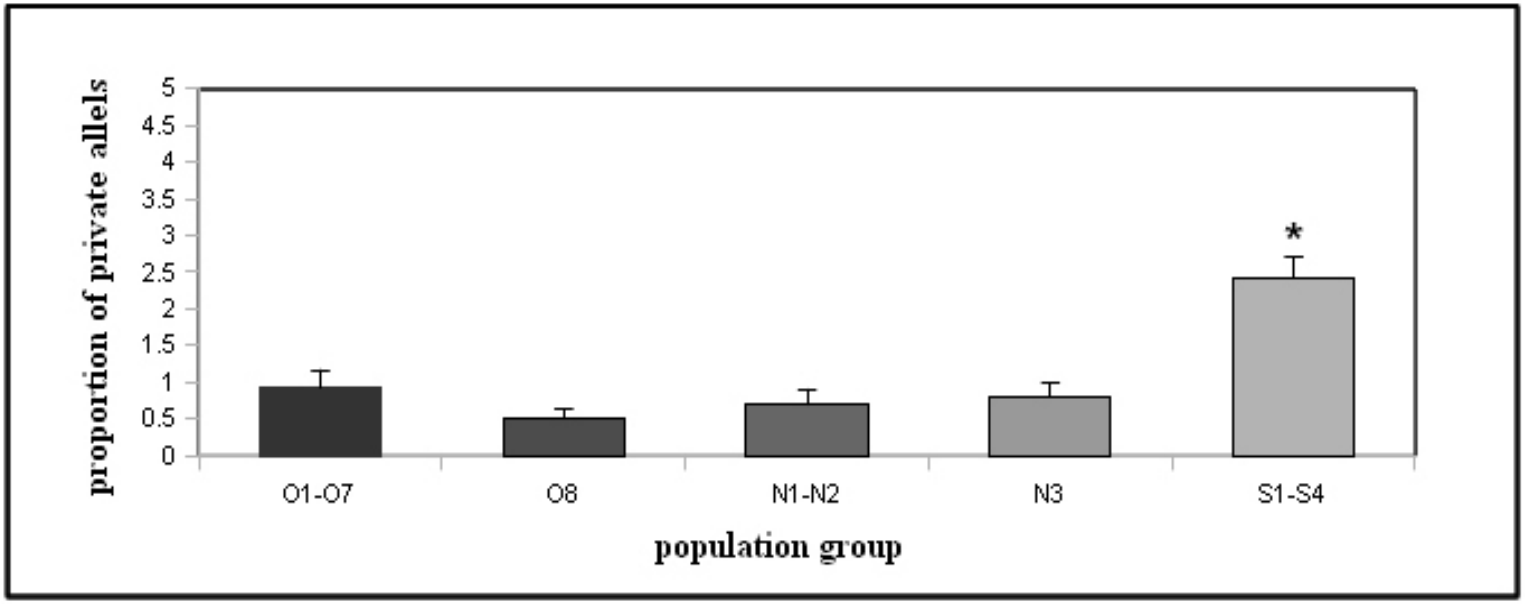



Figure 2.3. Genetic variability in *Astyanax mexicanus* using 26 microsatellite loci. A. Proportion of shared alleles (samples of likely common ancestry determined by shared alleles) between the studied populations shown as Euclidian distances, 95% confidence ellipses represent each population. B. Private allelic richness averaged over geographically grouped populations. Populations are coded as follows: El Abra caves (O1 - O7); Guatemala (N1 - N2); Micos (N3), Chica (O8); Surface (S1 - S4). All bar plots represent mean ± SEM. Asterisk denotes that the surface group was significantly different than each of the other groupings at the $P < 0.0001$, as tested by Student's t.

In order to determine genetic structuring in the analyzed samples, we performed hierarchical AMOVA analysis (Table 2.2). First, we narrowed down the population structuring by grouping populations based on their origins, “old” *vs.* “new.” Comparison of the El Abra populations (O1 - O8) *vs.* Guatemala (N1 – N2) and Micos (N3) caves pooled with surface populations (S1 – S4) was significant (P<0.0001) and explained 4.52 % of the variance among groups. This supports the hypothesis that two main stocks of surface fish were ancestral to the present day cave populations, as seen in the STRUCTURE analysis. Comparing “old” *vs.* “new” the analysis explained 3.4% of the variance (P<0.0001). Finally, the test of “new” caves *vs.* the surface populations also revealed significant differences (P < 0.0001) and explained 4.51% of variance among groups. The largest proportion of variance in all of the groups was within individuals (Table 2.2). We also tested the possible substructure of the surface populations grouped by geographical distance: El Abra (S3); Guatemala (S1); west of El Abra (S2); Micos region (S4), however, there were no significant differences (data no shown). AMOVA analysis supported significant groupings of three metapopulations: El Abra cave populations, Guatemala with Micos, and surface populations.

Pairwise $F_{ST}$ comparisons of the geographically defined populations typically revealed higher divergences among cave populations, even within a geographical cluster, than among cave and surface populations (Table 2.3). $F_{ST}$ comparisons revealed less divergence among populations of the two Guatemala caves (N1, N2) and Micos cave (N3) ($F_{ST}$ range from 0.23 to 0.36) than was seen in comparisons among caves of the El Abra cluster ($F_{ST}$ range from 0.20 to 0.51), even though the Sierra de Guatemala caves and Micos are geographically more than 100 km apart. $F_{ST}$ values among surface populations did not show big divergences (the highest $F_{ST}$ = 0.09) suggesting that many of these populations from multiple and distant geographical regions essentially have high levels of allelic exchange. On the basis of $F_{ST}$ values, general

divergence between cave and surface pairs seems to be related to the level of the geographical isolation of the particular caves from the surface water. Two El Abra populations (O6 and O1) as well as two Guatemala populations (N1 and N2) show the highest $F_{ST}$ values against the surface populations (Table 2.2.B). The first three of these populations are perched and thus isolated from the underlying aquifer, while the fourth is in an area with no permanent surface streams (Figure S2.1, Supplementary material).

The $F_{ST}$ analysis show statistically significant divergence between the O8 population and every other population of cave or surface fish we surveyed. The average $F_{ST}$ value between O8 and the seven other cave populations of the El Abra group (O1 to O8) was 0.230 $\pm$ 0.021 (SEM), which was significantly higher than the average $F_{ST}$ values between O8 and the ten surface populations (average $F_{ST}$ = 0.166 $\pm$ 0.006; $t_{11}$ = 5.75, $p < 0.0005$).

This single El Abra population (O8), while clearly aligned with the other old cave populations is much diverged from them. As is the case with O1 all seven $F_{ST}$ values between O8 and the other old cave populations are significant. With the exception of these two caves, the $F_{ST}$ values among El Abra caves are generally much lower (average $F_{ST}$ = 0.136 $\pm$ 0.017 *vs.* 0.230 $\pm$ 0.021 for O8 contrasts and 0.311 $\pm$ 0.010 for O1 contrasts) and the majority of them (10/15) are not significant.

| Structure tested | SS | VC | %VAR | Fstat | P |
|---|---|---|---|---|---|
| **1. O1-O8 *vs.* S1-S4 + N1-N3** | | | | | |
| Among groups | 372.71 | 0.43 | 4.52 | 0.07557 | **<0.000001** |
| Among populations within groups | 1288.50 | 1.51 | 15.76 | 0.16507 | **<0.000001** |
| Among individuals within populations | 3743.17 | 0.58 | 6.02 | 0.04517 | **<0.000001** |
| Within individuals | 3373.50 | 7.06 | 73.70 | 0.26303 | **<0.000001** |
| **2.O1-O8 *vs.* N1-N3** | | | | | |
| Among groups | 388.09 | 0.34 | 3.40 | 0.08321 | **<0.000001** |
| Among populations within groups | 990.49 | 2.60 | 26.27 | 0.27196 | **<0.000001** |
| Among individuals within populations | 2139.13 | 0.58 | 5.85 | 0.03398 | **<0.05** |
| Within individuals | 1882.00 | 6.37 | 64.48 | 0.3398 | **<0.000001** |
| **3.N1-N3 *vs.* S1-S4** | | | | | |
| Among groups | 202.03 | 0.49 | 4.51 | 0.05 | **<0.000001** |
| Among populations within groups | 500.00 | 0.92 | 8.44 | 0.09 | **<0.000001** |
| Among individuals within populations | 2708.85 | 0.86 | 7.91 | 0.09 | **<0.000001** |
| Within individuals | 2381.50 | 8.65 | 79.14 | 0.21 | **<0.000001** |

Table 2.2. Analysis of molecular variance (AMOVA) in cave and surface populations for 26 microsatellite loci. Bold P numbers represent significant values. SS - Sum of squares; VC - Variance components; % VAR - Percentage of variation; Fstat = F-statistics; P = P values.

| | EL ABRA | | | | | | | | GUATEMALA | | MICOS | SURFACE STREAMS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | O4 | O6 | O3 | O5 | O7 | O2 | O1 | O8 | N2 | N1 | N3 | S3 | S2 | S1 | S3 | S2 | S3 | S3 | S3 | S1 |
| O6 | **0.06** | | | | | | | | | | | | | | | | | | | |
| O3 | **0.19** | **0.25** | | | | | | | | | | | | | | | | | | |
| O5 | 0.06 | 0.15 | 0.21 | | | | | | | | | | | | | | | | | |
| O7 | 0.08 | 0.12 | 0.15 | 0.08 | | | | | | | | | | | | | | | | |
| O2 | **0.14** | **0.23** | 0.05 | 0.16 | 0.11 | | | | | | | | | | | | | | | |
| O1 | **0.28** | **0.34** | **0.31** | **0.30** | **0.34** | **0.28** | | | | | | | | | | | | | | |
| O8 | **0.22** | **0.23** | **0.26** | **0.18** | **0.16** | **0.23** | **0.33** | | | | | | | | | | | | | |
| N2 | **0.35** | **0.36** | **0.33** | **0.33** | **0.31** | **0.33** | **0.41** | **0.29** | | | | | | | | | | | | |
| N1 | **0.47** | **0.51** | **0.44** | **0.47** | 0.46 | **0.46** | **0.48** | **0.37** | **0.36** | | | | | | | | | | | |
| N3 | **0.24** | **0.26** | **0.26** | **0.23** | **0.20** | **0.24** | **0.31** | **0.23** | **0.23** | **0.27** | | | | | | | | | | |
| S3 | **0.16** | **0.19** | **0.15** | **0.13** | **0.08** | **0.14** | **0.23** | **0.16** | **0.19** | **0.26** | **0.11** | | | | | | | | | |
| S2 | **0.21** | **0.26** | **0.19** | 0.20 | 0.09 | **0.18** | **0.30** | **0.18** | **0.20** | **0.33** | **0.12** | 0.00 | | | | | | | | |
| S1 | **0.23** | **0.30** | **0.25** | 0.20 | 0.13 | **0.22** | **0.31** | **0.20** | **0.26** | **0.39** | **0.19** | **0.05** | **0.07** | | | | | | | |
| S3 | **0.18** | **0.22** | **0.18** | 0.17 | 0.10 | **0.18** | **0.26** | **0.15** | **0.22** | **0.34** | **0.11** | 0.02 | 0.03 | **0.09** | | | | | | |
| S2 | **0.19** | **0.22** | **0.17** | **0.17** | **0.10** | **0.17** | **0.25** | **0.19** | **0.20** | **0.28** | **0.13** | **0.02** | 0.01 | **0.08** | **0.04** | | | | | |
| S3 | **0.17** | **0.21** | **0.17** | **0.15** | **0.10** | **0.15** | **0.24** | **0.15** | **0.19** | **0.25** | **0.08** | 0.01 | 0.02 | **0.06** | 0.02 | **0.04** | | | | |
| S3 | **0.17** | **0.21** | **0.16** | **0.14** | 0.09 | **0.14** | **0.25** | **0.14** | **0.19** | **0.28** | **0.11** | 0.00 | 0.01 | **0.05** | 0.02 | **0.03** | 0.01 | | | |
| S3 | **0.18** | **0.22** | **0.18** | **0.15** | 0.09 | **0.16** | **0.26** | **0.16** | **0.19** | **0.28** | **0.13** | 0.00 | 0.02 | **0.03** | **0.03** | **0.04** | 0.01 | 0.00 | | |
| S1 | **0.18** | **0.22** | **0.18** | **0.16** | **0.11** | **0.16** | **0.25** | **0.16** | **0.20** | **0.28** | **0.13** | **0.02** | **0.03** | **0.05** | **0.03** | **0.05** | **0.02** | 0.01 | 0.01 | |
| S4 | **0.19** | **0.22** | **0.18** | **0.16** | 0.11 | **0.17** | **0.26** | **0.17** | **0.19** | **0.26** | **0.07** | **0.03** | **0.04** | **0.09** | **0.04** | **0.04** | **0.02** | **0.03** | **0.04** | **0.06** |

Table 2.3. Multilocus pairwise $F_{ST}$ estimates from 26 microsatellite loci in *Astyanax mexicanus*. Bold P values are significant values after Bonferroni correction.

**Effective population size and migration rates in *Astyanax mexicanus***

Estimations of effective population sizes ($N_e$) and migration rates among populations were performed with MIGRATE-N, using Bayesian inference and the Brownian motion mutation model. The model allows for mutation rates differing among loci by using the number of alleles per locus to estimate locus specific relative mutation rate modifiers. All the estimates of the mutation-scaled effective population size Θ were scaled using a microsatellite mutation rate of $5.56\times10^{-4}$ per locus per generation [196, 197] to calculate the average effective population sizes ($N_e$). Effective population size was variable between different surface clusters ($N_e$ from ~1011 to ~5058) but generally greater than in cave populations (Figure 2.4). Estimates of $N_e$ in most cave populations ranged from 831 (O6) to 1326 (O2) (Figure 2.4). However, the cave populations from which previous studies reported mixed populations were again an exception, with effective population sizes of 4159 in O8, 1326 in O2, and 2360 in the cave from the Micos region (N3).

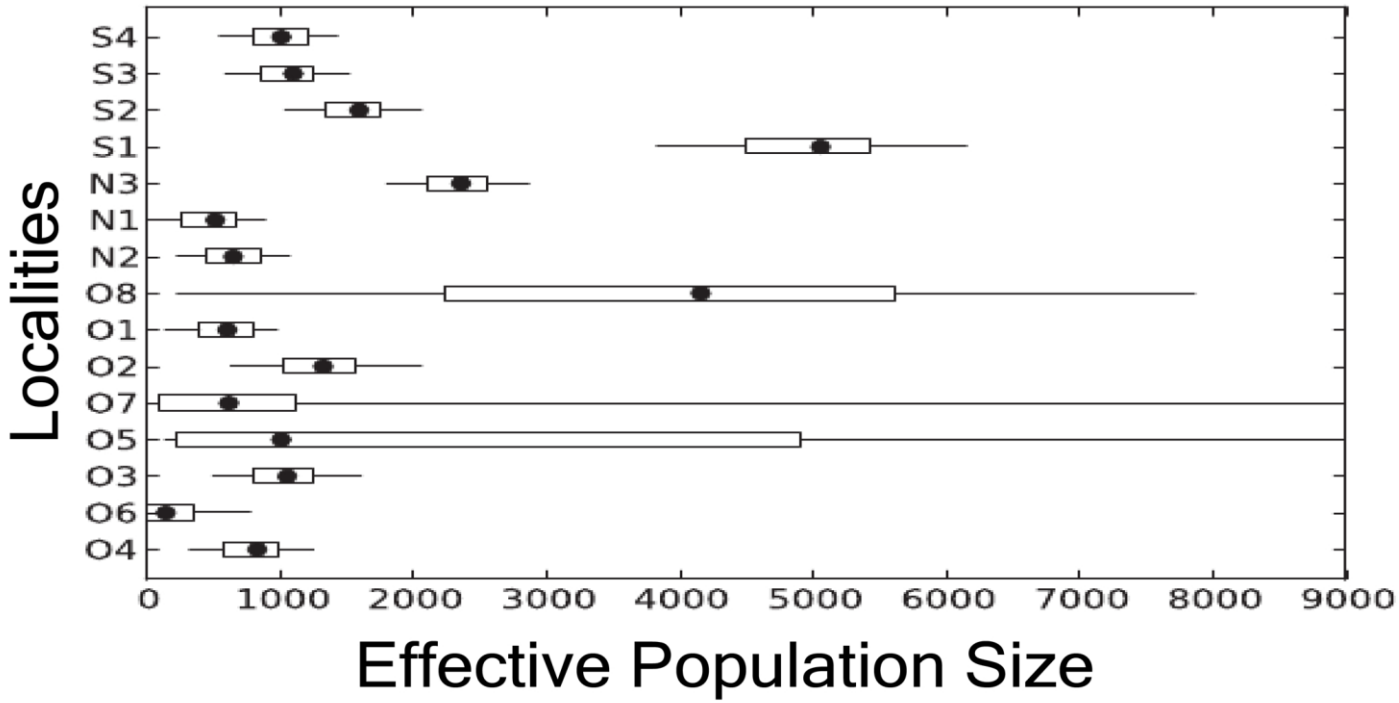



Figure 2.4. Estimates of effective population size ($N_e$) based on Bayesian inferences of migration rates and population sizes among *Astyanax mexicanus* population. The central box of the plots represents the values from the lower to upper quartile (25 to 75 percentile). The middle dot represents the median posterior values over all loci. The horizontal line extends from the 2.5% percentile to the 97.5% percentile. The x-axis represents $N_e$. Populations are coded as follows: El Abra caves (O1 - O7); Guatemala (N1 - N2); Micos (N3), Chica (O8); Surface (S1 – S4).

We used the MIGRATE-N models [198] to test for gene flow among individual cave and surface populations, limiting our inquiry to nearby populations or adjacent cave clusters. The summaries of all the models are presented in Figure 2.5 (see for the details Figure S2.2, Supplementary material). Our estimations of migration rates and effective population sizes supported the hypothesis that the genetic diversity of *A. mexicanus* cave populations is function of introgression from surface populations, as well as by the effective sizes of the cave populations.

Migration rates between individual populations varied by several orders of magnitude and the rates between cave and surface populations exceed those between caves. This is in accord with calculated $F_{ST}$ values. Four different patterns of migration were observed: among surface populations, among cave populations, from cave to surface, from surface to cave. Migration rates among the four groups of surface populations defined earlier (S1 - S4) were the highest we observed and were mostly symmetrical (Figure S2.2, Supplementary material). Migration rates between cave and surface populations were largely asymmetrical, with migration from the surface into caves typically greater than in the reverse direction. Micos (N3) cave was the only population that had almost the same migration rate in both directions, a result consistent with the STRUCTURE results. Migration rates among the cave populations were very low, except for caves that are geographically very close to one another in the El Abra cluster (O2, O3, O6, O4) or in the Guatemala cluster (N1, N2). Also, the new cave invasions (the Micos and Guatemala cave populations) seem to have more exchange of migrants with surface than with populations of the old cave cluster (Figure 2.5, Figure S2.2, Supplementary material). This suggests that proximate El Abra caves can exchange alleles through migration, although not nearly to the same extent as the surface populations exchange alleles.

Considering only the El Abra cave populations, we see that migration

rates decrease with increasing geographical distances among populations (Figure 2.5; Figure S2.1, Supplementary material). This observation supports the hypothesis of underground connections between nearby populations. Thus, O1 as the most geographically distant cave has the smallest influx from other cave populations of El Abra cluster, while O2, O3, O4 and O5 show high gene flow in both directions (Figure 2.5, Figure S2.2, Supplementary material).

In some cases the estimates of gene flow between two caves or cave clusters appear asymmetric. Considering both the Sierra de El Abra and the Guatemala, these asymmetries seemed related to relative altitudes. Figure 2.5 shows the altitudes above sea level of the fish pools in the various caves; N2 (175 m) sent more migrants to N1 (125 m) than vice versa. The same is true for O1 (202 m) to O2/O3 (147/153 m), O2/O3 to O4/O5 (62/84 m) and O7 (88 m) to O4/O5. Thus we suggest a gravitational model for gene flow. As water flows downward, so do alleles.

It must be noted that many of the estimates of migration rates are associated with large error terms (see details in Figure S2.2, Supplementary material) and are not precise. Nevertheless, the overall trends discussed above seem clear.

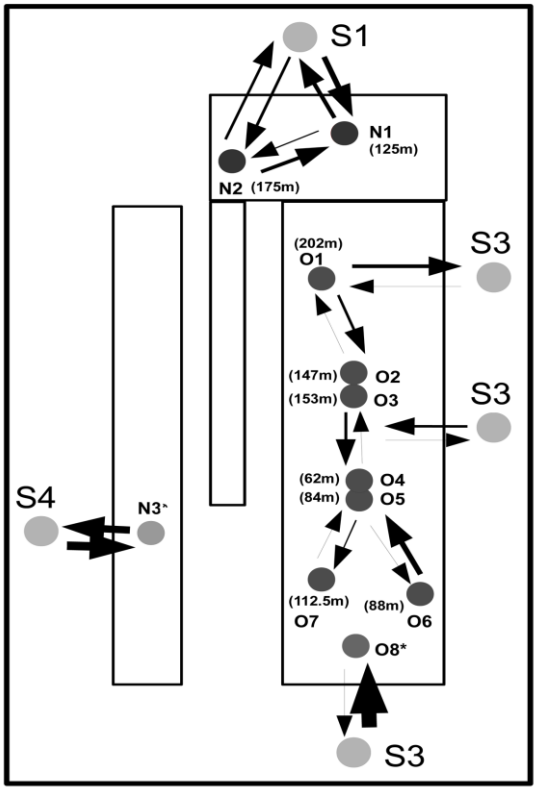



Figure 2.5. Summary of the estimates of gene flow based on Bayesian inferences of migration rates and population sizes using MIGRATE-N among *Astyanaxs mexicanus* population clusters within each geographical region. The arrows represent directions of migration and the thicknesses are proportional to the M (the ratio of immigration rate and mutation rate). Populations are coded as follows: El Abra caves (O1 - O7); Guatemala (N1 - N2); Micos (N3*), Chica (O8*); Surface (S1 - S4). Asterisk denotes mixed populations.

## Relationship between eye phenotype and individual admixture proportions

In order to understand the integration of the surface individuals into the cave in our populations we compared phenotype and genotype for individuals collected from the three caves with mixed populations. The phenotype we used was relative eye size and the genotypic designations for each of the 26 loci were obtained from the STRUCTURE analyses (Figure 2.6). Our results largely represent sorting of the phenotype and genotype into the two main categories,

surface and cave. In addition, however, we also observe that there are individuals that are in intermediate states in both genotype and phenotype, evidently hybrids between surface and cave.

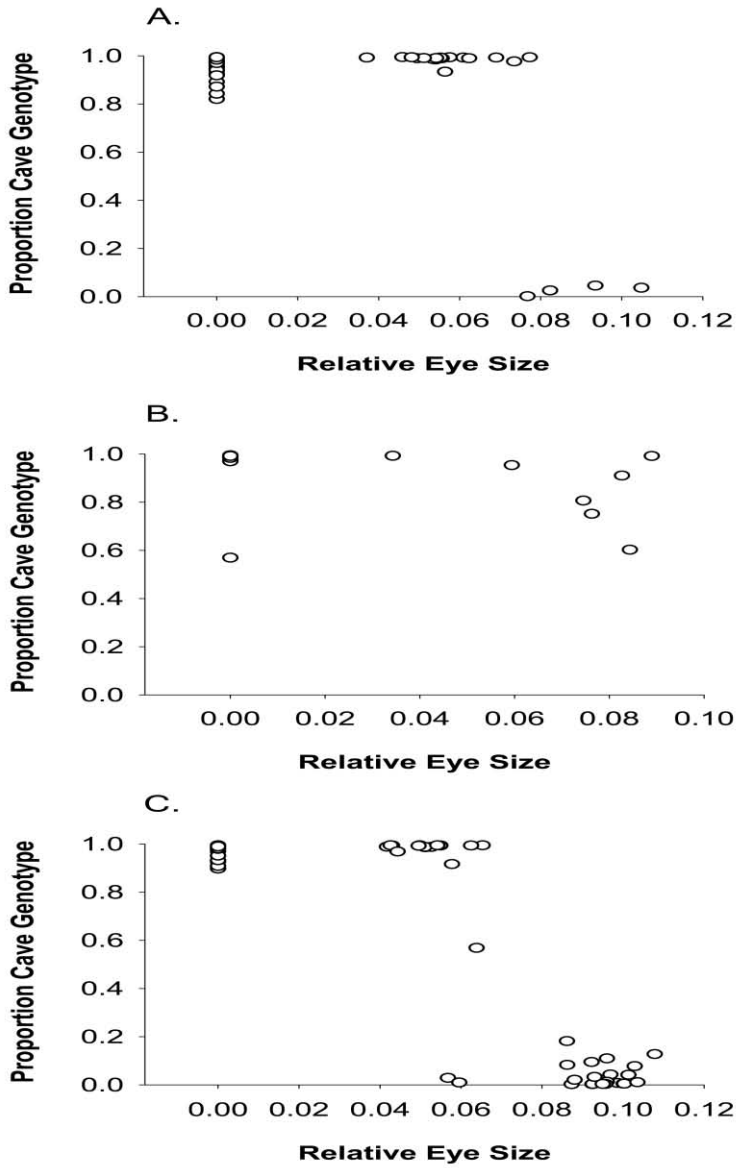



Figure 2.6. Relationship between genotype and phenotype in three mixed cavefish populations. Each point represents an individual fish. Phenotype is represented by relative eye size and genotype as the admixture proportions from the STRUCTURE analysis. A represents O8, B represents N2, C represents Micos.

## 2.4. DISCUSSION

### The Origins of the Cave Populations

Our data clearly show that the populations of cave adapted *Astyanax* in NE Mexico are derived from two separate stocks. Previous studies using microsatellites and mtDNA markers had also concluded that the cave populations were derived from at least two surface stocks [71-74]. Our results, however, show the affinities of the Pachón (O1) and Chica (O8) cave populations. Pachón was placed with the new stock based on mtDNA data, but our extensive nuclear DNA data set clearly places it with the old stock. The affinities of the Chica population are discussed below. Finally, the present study covers the full geographic range of the cave populations and reveals no evidence that the cave populations are derived from more than two clades.

Although derived from only two separate stocks, there are clearly more than two subterranean invasions that established the extant cave populations. All of our structuring analyses support divergence among five groups and are in accord with the hypothesis that the cave populations of Micos and the Sierra de Guatemala were established much later than the El Abra populations. Similarities in the microsatellite allele frequencies in the "new" cave populations (Molino, Caballo Moro, and Micos, N1 to N3, in order) and surface populations also confirm that these populations have recently diverged. With the exceptions of Pachón and Chica, the shared allele analysis shows that the El Abra populations cluster tightly. In the case of Pachón the divergence is minor and it is much closer to the old (El Abra) cluster than to the new cave cluster. In contrast, the Chica population is not obviously aligned with either cluster in the shared allele analysis.

The origin of the Chica population has been a long-standing question in

the *Astyanax* literature [65]. Our data strongly suggest that the Chica cave originated from old stock. This interpretation contrasts with a previous one based on mtDNA and a small number of microsatellite loci which suggested that it is phylogenetically young and originated from new stock [73]. If Chica were phylogenetically young, however, the STRUCTURE analysis should cluster it with the surface populations, a result not observed. Furthermore, we should see lower $F_{ST}$ values between Chica and the other "new" cave populations than between Chica and the other El Abra populations, but the opposite is the case (Chica vs. "new": average $F_{ST}$ = 0.297 $\pm$ 0.041 SEM; Chica vs. other El Abra: average $F_{ST}$ = 0.230 $\pm$ 0.021) (Table 2.3). Considering the $F_{ST}$ values, the STRUCTURE analysis, and the shared allele distance analysis (Table 2.3, Figure 2.2 and Figure 2.3A), all of which show Chica to be considerably differentiated from the rest of the El Abra populations, we suggest that it was derived from an independent invasion of old stock. Because of its southernmost location, it may well be the earliest established of the cave populations.

**Geology and Geography**

Knowledge of geology and geography, as well as genetics, is needed to understand the pattern of independent invasions of the underground that established the extant populations. Unfortunately, while we know the current geography well, we do not have a clear idea of how well the present state reflects the past. A clear pathway through surface waters from the southernmost end of the El Abra all the way to the area of Pachon cave existed in the past but at present a surface divide separates the ends of the valley [65] (Figure S2.1, Supplementary material). Pachón cave at the northern end of the El Abra is 46 km north of Yerbaniz cave (O2). While there is at least one other known cave between the two that might have served as a stepping stone, it seems likely that the underground invasion that established the Pachón cave

population was independent of those that established the more southern populations. This argument is based on the expectation that travel from one region to another is much faster through surface streams than through subterranean passages because open waters contain abundant food and provide direct passage, while subterranean routes have low food reserves and their passages may be maze-like. Surface fish can move into caves relatively easily and quickly. We constantly see surface *Astyanax* and other surface species, including *Tilapia*, in certain caves, such as Yerbaniz, Chica and Micos. The significance of Tilapia's presence is that it was introduced into Mexican waters and only became common in the late 1980's [199]. Therefore, its presence in caves shows how quickly underground populations may be seeded from the surface. Thus, for the most distal populations of a migratory wave, it is far likelier that surface migrants will have reached and colonized a cave long before the arrival of underground migrants from the same source. All seven $F_{ST}$ values between Pachón and the other El Abra populations are significant (Table 2.3), which reflects the current isolation of the cave and, perhaps, a past independent origin.

Considering the new cave populations, the distance between the Micos cave and the closest of the Guatemala caves, Caballo Moro, is over 90 km and there is one ridge and two open valleys between them. No documented underground route currently exists between the two regions. Thus, the Micos and Guatemala cave clusters likely represent separate invasions.

In summary, we suggest a model with at least five independent origins of cave adapted *Astyanax* in NE Mexico (Figure S2.3, Supplementary material): 1) Chica in the south, 2) Pachón in the north, 3) the remaining El Abra populations, 4) the Sierra de Guatemala cave populations, and 5) the Micos cave populations.

**Allelic diversity, migration and gene flow**

Allelic diversity, by all measures, was generally lower in cave populations than in surface populations (Table 2.1), an observation in accord with previous studies on cavefishes and other cave organisms [63, 71, 73, 200]. Lower genetic diversity in cave populations than in related surface populations probably reflects smaller effective population sizes because of food and space limitations, but may also reflect possible bottleneck events due to periodic droughts and other environmental fluctuations [65]. It should be noted, however, that the relatively large effective population sizes in Micos and Chica were probably overestimated by MIGRATE-N because they are admixed with the surface.

Many of the El Abra caves regularly receive migrants from the surface [65, 200], and Chica is the best known of these [64]. Chica is unusual among *Astyanax* cave populations in receiving a high energy input deriving primarily from two bat roosts located directly above the largest of the fish pools. Breder noted, and we still observe today, that the frequency of surface fish in the pools increases as one goes deeper into the cave, and is highest in Pool 4, at the level of the aquifer and located about one km from the Río Tampaón [64]. All who have studied this cave have surmised that surface fish get into the cave from the river through the aquifer and are able to survive and breed there because of the high energy input from the bat roosts and from debris washed into the cave during the rainy season [64, 65]. Thus, Chica draws its occupants from two different source populations that are well differentiated from each other. This admixture results in significant heterozygote deficiencies at numerous loci. That these departures from HWE are due to Wahlund effect is evident from genotype-phenotype correlations observed in our study (Figure 2.6).

Our collections from the Micos cave also contained both cave and surface forms and, as in Chica, we observed departures from HWE, due to

Wahlund effects. In contrast to the situation in Chica, food is not abundant in this cave, thus the surface fish are prone to starvation, leading in most cases to reduced fitness and inefficient mating [65]. Nevertheless, some surface fish washed into this cave may hybridize with the cave population, as revealed by genotype-phenotype correlations (Figure 2.6). The Caballo Moro population exhibits a full range of eye sizes and pigmentation from typical cave to typical surface morphs (Figure 2.6). This population is in a karst window, a habitat within a cave exposed to light because of passage collapse; the presence of light facilitates the continued survival of surface and hybrid phenotypes [65, 201].

The MIGRATE-N analysis also detected relatively high rates of gene flow from the Pachón cave population to their nearby surface populations, supporting an earlier suggestion of a route for alleles from cave to surface [200] (Figure 2.5, Figure S2.2, Supplementary material). Estimation of migration rates and effective population size supported the hypothesis that the genetic diversity of *A. mexicanus* cave populations is correlated with the influx of alleles from surface populations, as well as by the effective population sizes [200].

The migration rate analysis revealed that surface fish in the region form a metapopulation, with extensive exchange of genetic material among its component populations. Thus, there is high genetic diversity within and little genetic differentiation among surface populations. In strong contrast, cave populations live under dramatically different ecological conditions and often have lower population densities. MIGRATE-N results also show that the effective sizes of surface populations are generally larger than those of cave populations, consistent with earlier studies based on estimates of nucleotide diversity [200] (Figure 2.4). Mark and recapture estimates of total population sizes from Pachón and Yerbaniz caves were similar, with averages of $8.5 \times 10^3$ individuals and broad 95% confidence intervals ranging from about $1.5 \times 10^3$ to

$17.0x10^3$ [65]. Our estimates of cave population $N_e$ varied from 2.8 to $7.3x10^3$ with the exception of Curva and the admixed populations (Micos and Chica) and are consistent with the estimates reported in Mitchell et al. [65], and around one order of magnitude higher than previously reported [200].

We note that the mutation-scaled immigration rate M from surface populations into cave populations often exceeds 1 (Figure S2.2, Supplementary material). With mutation-scaled effective population sizes Θ (Theta) on the order of 0.5 to 5, m x $N_e$ (Θ x M/4) can exceed 1.0, implying that migration from surface to cave populations could significantly affect allelic frequencies at neutral loci [96]. Nevertheless, cavefish in these populations remain troglomorphic in phenotype in the face of this immigration, which implies that these phenotypes are maintained by selection. Selection may generally be sufficiently powerful to allow population differentiation even in situation in which there is high gene flow [202].

Finally, we note that the five independent invasions of the subterranean habitat documented here imply five instances of striking phenotypic convergence. This highlights the importance of a change in ecology as a strong driver of evolutionary change. This is in accord with studies of freshwater adaptation in *Gasterosteus aculeatus* that document widespread convergences or parallelisms related to ecological shifts [5].

## 2.5. MATERIAL AND METHODS

### Sampling

All fish specimens were collected in March 2008 and preserved in 70% ethanol. A total of 568 *Astyanax* samples were taken from 11 cave and 10 surface locations. Names and abbreviations of the sampled populations, and other details are listed in Table 2.1. Samples collected from caves can be divided into three geographically distinct groups: the Sierra de El Abra cave

cluster, the Western slope of the Sierra de Colmena (= the Micos area), and the Sierra de Guatemala. The El Abra cave cluster is represented by eight caves (from North to South, O1 to O8): Pachón, Yerbaniz, Japones, Arroyo, Tinaja, Curva, Toro, and Chica, respectively. In the Sierra de Guatemala we sampled two caves, Molino (N1) and Caballo Moro (N2) and in the Micos area we sampled only one of three closely clustered caves (Río Subterráneo): Micos (N3). An overview of the geographical distribution of the sampling area of cave and surface locations is presented in Figure 2.1 and the locality abbreviations are shown in Table 2.1.

DNA extraction and genotyping were done according to Protas *et al*. and all samples were profiled at 26 microsatellite markers with primers previously developed for QTL studies [78]. We used unlinked markers selected from independent linkage groups, or so distant as to assort independently if within the same linkage group [77]. The forward primer of each pair was labeled at the 5`end with a fluorescent dye (HEX or FAM) and microsatellite amplification products were visualized on an ABI 3730 automated DNA sequencer. Microsatellite markers were optimized for the allelic range and multiplexed. Allele sizes were scored using GENEMAPPER v3.7 (ABI).

**Genetic diversity**

We calculated observed ($H_o$) and unbiased expected ($H_e$) heterozygosities [203], number of alleles, and the number of alleles standardized for the smallest sample size for single populations and for the geographic groups. These descriptive statistics were performed in Genepop v 4.0 [204] and Microsatellite Analyzer (MSA) [205]. Deviations from HWE were estimated using both the exact test and the $F_{IS}$ statistic estimations, using Markov chain Monte Carlo (MCMC) runs for 1000 batches, each of 2000 iterations, with the first 500 iterations discarded before sampling [206]. Whenever multiple testing was performed, probability values were corrected using standard Bonferroni corrections [207].

**Population structure analysis and differentiation**

The program STRUCTURE 2.3.3 [159] was used to infer historical lineages through clustering of similar genotypes. The admixture model of STRUCTURE and the option of correlated allele frequencies between populations were used. The correct number of clusters was determined by testing K values from 1 to 12 and performing 10 repeats for each K. The burn-in period consisted of 1 x $10^6$ iterations followed by 1 x $10^5$ MCMC repeats. Finally, estimated log probabilities of data Pr (X | K) for each value of K were evaluated by calculating ΔK, the rate of change in the log probability of data between successive K values [195]. In order to confirm the STRUCTURE inferences, population structure was additionally estimated using STRUCTURAMA [208].

While these clustering methods can be quite powerful, particularly when there is a high divergence between populations [209], they often make explicit assumptions of demographic history and sometimes are difficult to interpret without background biological information.

Thus, we complemented the Bayesian analysis using other methods to more directly estimate relationships among populations. The proportions of shared alleles between populations were calculated in the R package adegenet 1.2-2 using the *propShared* function [210], where the average proportions of shared alleles among and within populations are computed over all possible combinations of individuals sampled. The distance matrix based on the proportion of shared alleles was then transformed into a matrix of Euclidean distances using the *quasieuclid* function.

Private allele estimates and allele richness were calculated, grouping the independent geographical regions obtained by clustering methods. In order to estimate rarified allelic richness and private rarified allelic richness, the rarified method in HP-RARE [211] was used to control for the correlation between observed allelic diversity and sample size [212]. The alleles were

rarified to a sample size of 40, the smallest sample size of our population groups.

In order to estimate the variance between the groups of populations, pooled sample structuring was estimated using analysis of molecular variance (AMOVA) [185] and 20,000 permutations implemented in ARELQUIN v 3.5.1.2 [213]. Influences of long-term separation and genetic drift were measured by comparative methods of allelic frequency tests for all population combinations using $F_{ST}$ pairwise estimates [139] as implemented in MICROSATELLITE ANALYSER (MSA) [205].

**Migration patterns between populations**

The coalescence-based program MIGRATE-N [198, 214, 215] was used to test for and estimate gene flow between populations. Three migration models were evaluated: (1) a full model with two population sizes and two migration rates (in and out of the cave); (2) a model with two population sizes and one migration rate (gene flow into the cave); (3) a model with two population sizes and one migration rate (gene flow out of the cave).

MIGRATE-N also estimated the mutation-scaled effective population size (Θ) (i.e. 4 $N_e$ x μ, where $N_e$ is the effective population size and μ is the mutation rate per generation per locus) and the mutation-scaled migration rate M (M = m / μ, where m is the immigration rate per generation) among cave and surface *Astyanax* populations. The model comparison was done using Bayes factors that need the accurate calculation of marginal likelihoods. These likelihoods were calculated using thermodynamic integration in MIGRATE-N 3.1.7 [198].

*MIGRATE-N 3.1.3 runtime condition*

Most run parameters for the program MIGRATE-N 3.1.3 were left at default

values, but adjustments were made on parameters influencing the run-length, heating, and relative mutation rate, and, of course, to specify different migration models.

The mutation rate among loci was scaled so that the average rate change of the mutation rate was 1.0. This equalizes the effects of the estimates of the individual loci; relative rates changed for different runs because of data differences, but commonly the minimum (0.0369) and maximum rates (2.32) were rare and most datasets had ranges for the rate of mutation rate change of about 0.6 to 1.5.

Per locus the first 100,000 steps were discarded, then 2.5 million steps were visited using parallel runs of 100 replicates. These resulted in recorded 50,000 samples that were recorded every 50th step. A step comprises of either a parameter change or a genealogy change. A total of 26 loci yielded samples of 65 million steps. To improve searching and also to calculate marginal likelihoods for the model comparison a heating scheme was applied using 4 changes with temperatures 1.00, 1.50, 3.00, and 1000000.00.

A random genealogy and parameter settings inferred by an $F_{ST}$-based method where used as start condition. The prior distribution for the parameters was uniform with boundaries appropriate for the parameters and data: Theta priors were bounded between 0 and 50.0 and M priors where bounded between en 0.0 and 100.0.

**Eye size measurements**

For the phenotypic comparison between cave and surface-dwelling specimens, we analyzed digital images of individuals from three cave populations (N2 (n = 26), N3 (n = 72) and O8 (n = 119)). Photos were taken in the lab using a digital camera with the fish placed on a Cartesian coordinate grid. Measurements were made using ImageJ (NIH). In order to correct for individual size

differences, relative eye size was standardized as a proportion of standard body length [66].

## 2.6. ACKNOWLEDGMENTS

Martina Bradic design the project and carried out all of the experiments in this section, preformed the data analysis and wrote the paper. P.B. helped with Migrate-N analysis and data interpretation. R.B supervised the project, helped with writing the paper and discussion of the data. F.J.G.L. and S.C provided the collection permit and helped with samples collection.

Microsatellite primers were provided from J. Gross (Harvard University Medical School). We thank the Mexican government for providing the collecting permit (DGOPA.00570.288108-0291). This work was funded by a Fundação para a Ciência ea Tecnologia PhD grant to M.B (SFRH/BD/32982/2006), and an NSF IOS - 0821939 awards to R.B. P.B. was partly funded by the joint NSF/NIGMS Mathematical Biology program under NIH grant R01 GM 078985 and by NSF grant DEB 0822626. F.J.G.L. was partly founded by the CIBNOR.We thank Paul Scheid for technical help, Erik Duboué and the members of Teotónio laboratory at Instituto Gulbenkian de Ciência, Portugal for the useful comments on the manuscript.

# CHAPTER 3

## Signatures of selection on standing genetic variation and association with adaptive phenotypes in the cave environment

Manuscript in preparation

## 3.1. SUMMARY

Instances in which we observe a repeated phenotypic occurrences in the same type of environment allows for direct testing of the natural selection. Using a data set of 745 genome-wide SNP markers we investigated the effect of natural selection on the maintenance of differentiation in multiple cave-surface fish comparisons. To associate detected SNPs with the phenotype we used $F_2$ cross between the cave and surface individuals and genotyped the same SNP markers in $F_2$ progeny. We have designed a genetic map and performed QTL analysis for ten phenotypes. Further, we detected SNPs in multiple natural populations also falling into QTL loci for lens, amino-acid sensitivity and eye size. We observed haplotypes that were repeatedly selected in cave populations of the new lineage but were present in very low frequencies in the surface populations, or at such low frequencies as to elude detection. These suggest that adaptation from standing genetic variation plays an important role in the adaptation to the cave environment. We also observed that parallel genetic evolution occurs more frequently among closely related populations, suggesting the importance of evolutionary history in parallel and convergent genetic change. Furthermore, we observed the alternative possibility that implies that natural selection can repeatedly generate similar patterns of genotypic variation in different ways (different haplotypes within each lineage). Finally, convergent phenotypic change in different populations can arise through a conserved genetic basis.

## 3.2. BACKGROUND

Disentangling the genetic basis of adaptive phenotypic variation is central to our understanding of the origins and maintenance of biological diversity. However, identifying the genetic changes responsible for phenotypic adaptation is extremely difficult and to date, successful attempts have been limited to a selected group of model systems, such as fruit flies (e.g. [22, 56, 112, 114, 216], worms [217-219], stickleback fishes [4, 5, 7-9, 169, 220, 221] and mice [113, 115, 192, 222-224].

Studies of repeated phenotypic changes in the wild provide an excellent laboratory to test natural selection on morphological trait that evolves multiple times. Thus, this gives us a very powerful tool to test if particular morphological changes are adaptive. Furthermore we can ask how those changes vary between distantly and closely relate populations and are the same loci involved in the adaptive evolution of similar traits.

New era in genomics is giving us a possibility to perform genome-wide studies of morphologically similar replicates and test for above-mentioned possibilities [119, 120, 155, 191, 225-227]. Genetic diversity between populations is frequently used to identify the specific genomic regions that exhibit significantly increased or decreased differentiation among populations in order to identify candidate loci [131, 154, 156, 157, 163, 165, 228]. Differentiation among populations is a function of the number of migrants per generations ($N_e$ x m). When population sizes are large, even low migration rates among populations can prevent differentiation at neutral loci, although not at adaptive loci [157]. Migration at neutral loci will homogenize allelic frequencies between two populations even if there is as little as 1 migrant per generation [96]. So, overall we will not see divergence between the populations at those loci. In contrast, for the loci that are related with the fitness to the certain environment, migration will be counteracted by local adaptation. Thus, by contrasting patterns of diversity at numerous loci across

the genome it is possible to find those specific regions ("outliers") that have likely been under strong diversifying or stabilizing natural selection rather than under neutral genome-wide effects (genetic drift, migration and inbreeding) [154, 156, 157, 163, 229, 230]. Several methods have been developed to identify regions of genetic divergence based on this hypothesis (e.g. $F_{ST}$ outlier genome scan methods [154, 156, 163, 166, 185, 213]). Nevertheless, any of these approaches has its limitations and may be biased towards identifying only markers under particularly strong selective pressure [57]. Thus, genome scan methods should be complemented by other approaches towards linking the effect of selection with genetic and, ultimately, adaptive phenotypic divergence [8, 106, 155, 157, 231]. An integrative approach using the above-mentioned methods for selection detection together with traditional QTL mapping could provide additional information on the alleles or loci that co-segregate with certain QTL [5, 8, 32, 44, 46, 105, 106, 119, 231, 232].

The *Astyanax mexicanus* system with five independent invasions of the subterranean habitat implies five instances of striking phenotypic convergence and is well suited to answer the above-mentioned questions (Chapter 2). Based on its evolutionary history this species offers a unique opportunity to investigate whether the evolution of parallel and/or convergent phenotypes occur thorough changes in the same genetic loci. Because our goal is to make a connection between evolution at the phenotypic and genotypic levels, we need a clear definition that bridges the phenotypic and molecular level. Thus, we define parallel genotypic adaptation as the independent evolution of homologous loci to fulfill the same function in two lineages. Based on this definition, changes at non-homologous loci resulting in the same phenotype are considered convergent (according to [2]).

In order to identify adaptive loci in natural populations of *Astyanax mexicanus* and detect if those adaptations are common in the genome, we used a genome-wide scan of multiple SNP markers developed by next

generation methods (RAD-tag sequencing) in natural populations and in laboratory crosses. We contrasted multiple markers across the genome and across multiple populations in order to distinguish adaptive loci that were further associated with the quantitative trait loci (QTL) available from genetic cross.

Due to the known ancestral state (surface fish) we also ask if the adaptations to the novel environment are the result of new mutations or preexisting genetic variation in ancestral population [5, 54, 58, 59]. We addressed this question by comparing the ancestral allele state and alleles of the multiple independent populations across identified QTL regions. In this study we applied an integrative approach that combines differentiation within and among the independent populations of *A. mexicanus* with classical QTL mapping. We detected loci that are repeatedly segregating in closely related cave populations strongly suggesting biological significance and evolutionary history related to local adaptation in the cave.

## 3.3. RESULTS

### *De novo* sequencing and SNP discovery

The individuals sequenced for SNP discovery were three F1 heterozygotes from a cave x surface cross. Sequences were scanned to identify heterozygous SNPs present in all three individuals (i.e. both alleles are present in fish reflecting a heterozygous state). We obtained equal amounts of DNA from all three individuals using RAD tag technology and the sample normalization was within expectations, with approximately 1.5 million reads obtained per sample [233, 234]. On average each tag was sequenced approximately six times in every individual. This depth of coverage allowed the identification of SNPs that were further verified in genotyping by Sequenom; sites at which coverage was insufficient (lower than 6x) were not considered. General metrics for the assembly are detailed in Figure 3.1. We recovered

43,282 contigs with an average contig length of 230.9 bp and a total of 9.993 Mb *de novo* sequence, 0.9% of the 1.2 Gb *Astyanax* genome. Within this sequence, 9900 putative SNPs and 1000 indels were identified. We chose 556 RAD SNP markers suitable for downstream genotyping design and those contigs were BLAST-ed to the Zv8 *Danio rerio* genome assembly in order to determine possible gene functions. Fewer than 50 contigs anchored to established zebrafish positions using moderate stringency alignment (bit score > 50) (due to a large amount of data those sequences are not present in this thesis and will be only submitted to the database once manuscript submitted).

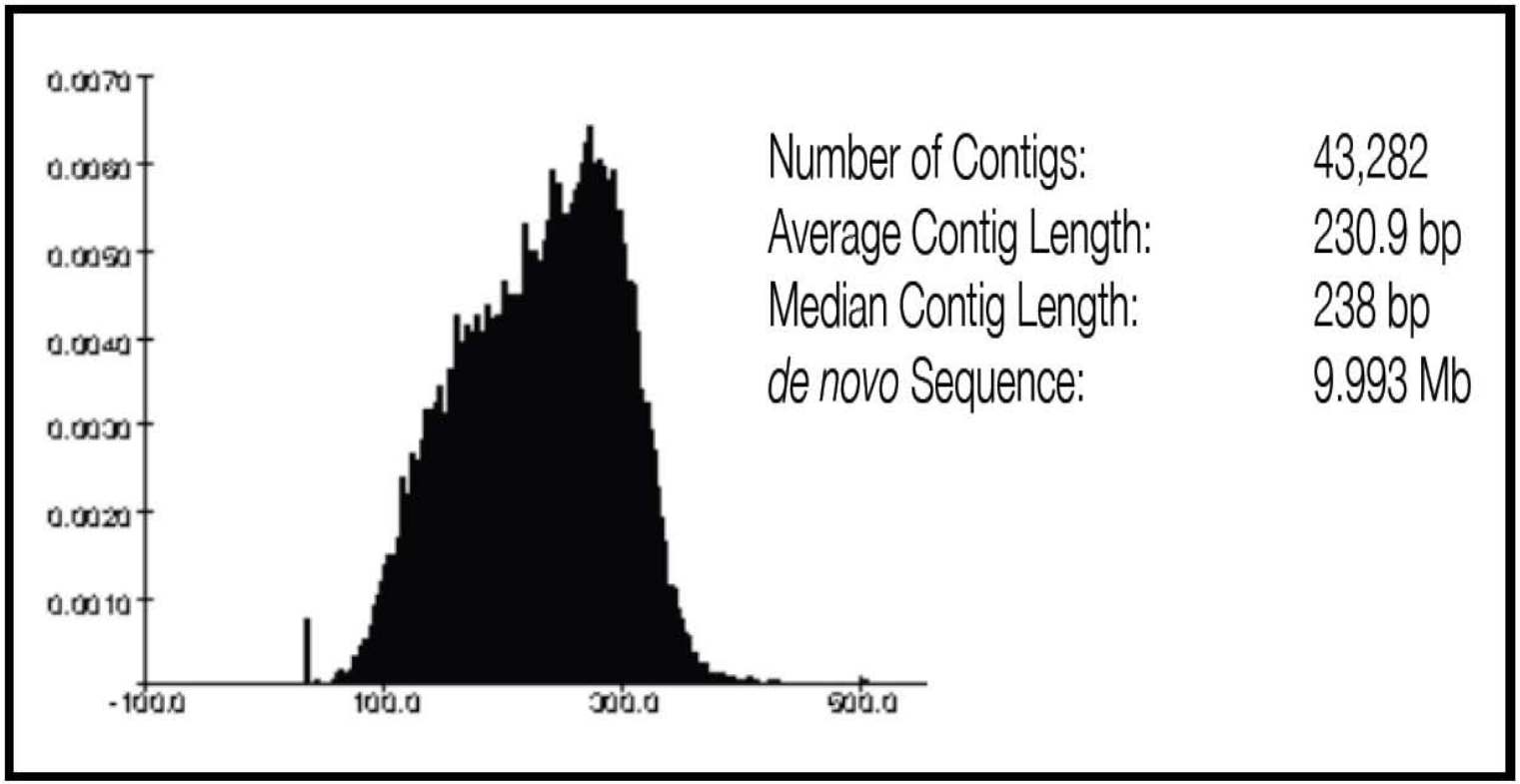



Figure 3.1. Summary of Astyanax contig statistics using RAD tag sequencing methodology (see details in the text). Distributions of the contig lengths are shown together with the contig statistics.

*Illumina paired-end BAC sequencing and SNP discovery*

The paired-end data from this Illumina run suffered from low sequence quality due to the high number of repeats that made it challenging to interpret and include in the assembly. Nevertheless, we were able to successfully piece together large portions of each BAC. We have considered sequences of greater than 500 bp with minimum sequence coverage of 6x over the length of the contig. The median contig lengths (N50) of the BACs indicate that the BAC containing the growth hormone (BACGH) gene was the easiest to assemble (likely originating in a region with little repetitive or low complexity sequence) while BAC6 and BAC10 were more fragmented. N50 values in BACGH, BAC1, BAC6, and BAC10 were 11.3 kb, 10 Kb, 7.75 kb and 3.65 kb respectively. Total assembled sequences in those three clones were 122 kb, 100kb, 79 kb and 82 kb respectively. Additional short sequence fragments (~3kb in average) produced by short read (35bp) Illumina trial runs were also used in SNP discovery. We have detected a total of 188 SNP markers from 7 BACs and candidate genes used in Sanger sequencing, which were further used in genotyping.

**Linkage map and QTL loci**

After quality control and removal of low quality data as described in methods, the data set contained 474 high quality SNP genotypes for 273 individuals from the $F_2$ mapping cross. The same individuals had 259 microsatellite markers available from the previous mapping project [77]. We have constructed two linkage maps using two types of markers: 1) An integrated map that includes both SNP and microsatellite markers [77] and 2) a SNP map that includes only the SNP markers developed in this study as well as SNP markers previously typed for the candidate genes in this cross as described in Protas et al. [77].

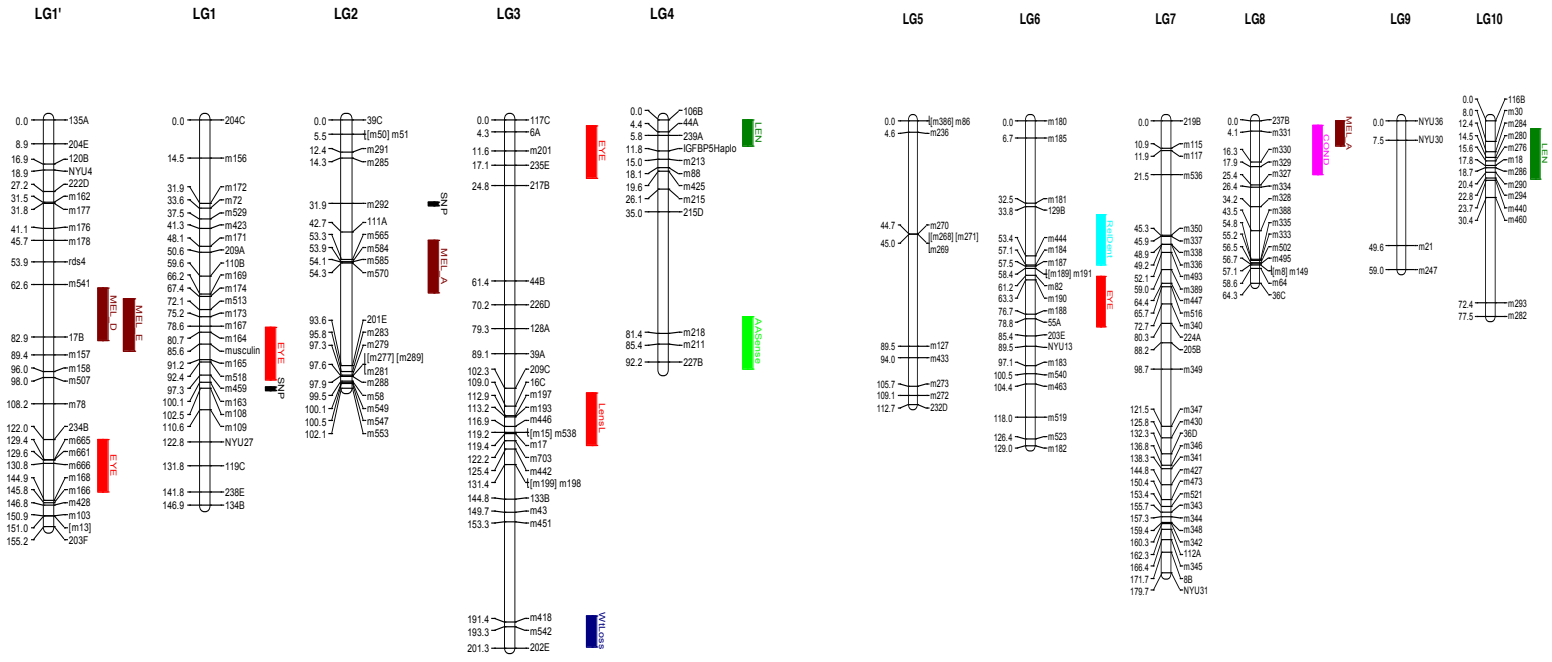
The integrated linkage map contained 40 linkage groups and 450 markers. The length of the map was 2646.25 cM with an average interval

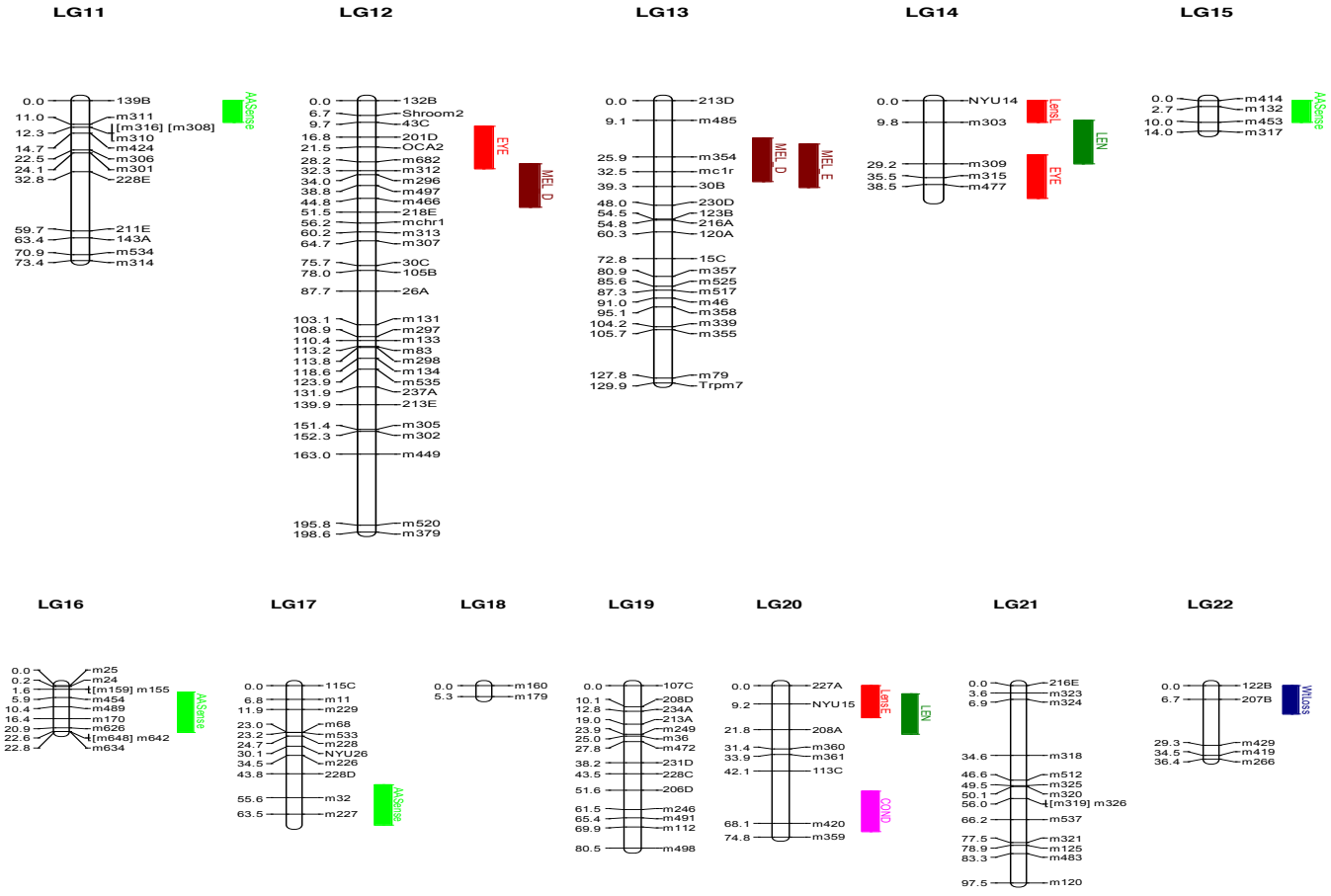
length between markers of 6.46 cM (Figure 3.2). In order to retain the information about the relative positions between the SNP markers we also constructed a second map that contained only SNPs. This map consisted of 24 linkage groups and 374 markers and was shorter in length than previously observed (1904 cM *vs.* 2148cM) [77] and the average interval lengths was 5.44 cM (Figure S3.1, Supplementary material). These results suggest that linkage map containing only SNP markers is more informative than the microsatellite map, which is probably due to better reliability of the SNP markers.

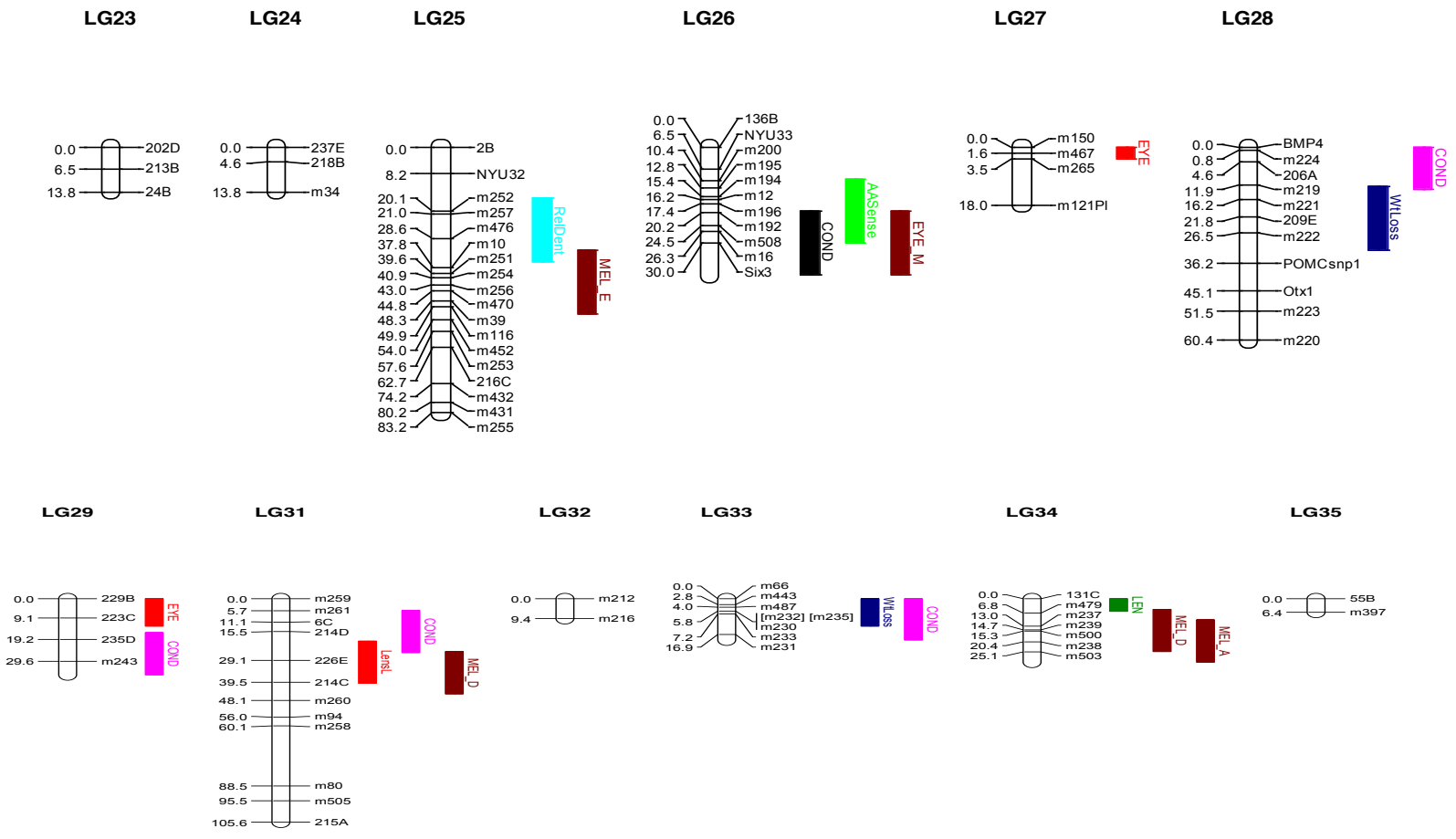
To assess the effectiveness and accuracy of mapping the QTL traits we used the integrated map and repeated the QTL mapping using already available phenotypes [77]. A detailed description of each phenotype is given in Table 3.1.A. We re-mapped ten traits that differ between cave and surface tetras and detected 50 QTLs (Figure 3.2, Table 3.1.B). The number of QTL per trait ranged from one for estimated daily growth rate (GrLen) to eight for relative eye size (RelEye) (Table 3.1.B). The position of the maximum LOD score as the best estimate of QTL position is shown in figure 3.2 as colored for the individual trait and the QTL width is defined by maximum LOD +/- 1.0 LOD. Detailed LOD profiles for the specific traits and linkage groups (LG) were explored more in detail, later in the chapter (*see haplotype phasing and diversity*).

The percentage of total trait variance explained (PEV) per QTL ranged from 10 to 15 % and the additive trait variance (PEVad) per QTL was ~7%. The highest PEV and PEVad was observed for eye size (RelEye) in LG14, count of melanophores around the eye (EyesMel) in LG13, number of thoracic ribs (ribs) in LG40, count of dorsal melanophores (DorsalMel) in LG13 and amino-acid sensitivity (AAsense) in LG26 that accounted for 48%, 35%, 30%, 27% and 27% of PEV, respectively (Figure 3.3) (Table S3.1, Supplementary material).

We also observed clustering of QTL for different traits on the following linkage groups: condition factor, lens size and count of the number of melanophores in a defined dorsal region in LG31 (COND, LensL, MEL_D), count of the number of melanophors in a defined dorsal and lateral region and lens in LG34 (MEL_D, MEL_A, LEN), amino-acid sensitivity, eye melanin and condition factor in LG26 (AAsense, EYE_Mel, COND) and lens, length and eye size in LG14 (Lens, LEN, RelEye). Some of the QTLs even overlapped their LOD range (Figure 3.2). The smallest detected PEV was as low as ~ 1.5% and was detected in relative eye size (RelEye) QTL in LG3 (Figure 3.3; Table 3.2, Supplementary Material). In summary we have observed traits where QTL explained as much as ~50% of the trait variance (LG 14, eye QTL) to the loci that had up to 8 QTLs that explained small proportion of the trait variance. However, it might well be that the locus of adaptation is a single one, despite the functional differentiation of several QTLs which we further explore in genome-wide scan of the natural populations.


LG1'
LG1
LG2
LG3
LG4
LG5
LG6
LG7
LG8
LG9
LG10
MEL_D
MEL_E
EYE
SNP
MEL_A
LensL
WLloss
LEN
AASense
RetPart
COLOR


LG11
0.0 139B
11.0 m311
12.3 [m316] [m308]
m310
14.7 m424
22.5 m306
24.1 m301
32.8 228E
59.7 211E
63.4 143A
70.9 m534
73.4 m314
AASense
LG12
0.0 132B
6.7 Shroom2
9.7 43C
16.8 201D
21.5 CCA2
28.2 m682
32.3 m312
34.0 m296
38.8 m497
44.8 m466
51.5 218E
56.2 mchr1
60.2 m313
64.7 m307
75.7 30C
78.0 105B
87.7 26A
103.1 m131
108.9 m297
110.4 m133
113.2 m83
113.8 m298
118.6 m134
123.9 m535
131.9 237A
139.9 213E
151.4 m305
152.3 m302
163.0 m449
195.8 m520
198.6 m379
EYE
MEL D
LG13
0.0 213D
9.1 m485
25.9 m354
32.5 mc1r
39.3 30B
48.0 230D
54.5 123B
54.8 216A
60.3 120A
72.8 15C
80.9 m357
85.6 m525
87.3 m517
91.0 m46
95.1 m358
104.2 m339
105.7 m355
127.8 m79
129.9 Trpm7
MEL D
MEL E
LG14
0.0 NYU14
9.8 m303
29.2 m309
35.5 m315
38.5 m477
LensL
EYE
LEN
LG15
0.0 m414
2.7 m132
10.0 m453
14.0 m317
AASense
LG16
0.0 m25
0.2 m24
1.6 [m159] m155
5.9 m454
10.4 m489
16.4 m170
20.9 m626
22.6 [m648] m642
22.8 m634
AASense
LG17
0.0 115C
6.8 m11
11.9 m229
23.0 m68
23.2 m533
24.7 m228
30.1 NYU26
34.5 m226
43.8 228D
55.6 m32
63.5 m227
AASense
LG18
0.0 m160
5.3 m179
LG19
0.0 107C
10.1 208D
12.8 234A
19.0 213A
23.9 m249
25.0 m36
27.8 m472
38.2 231D
43.5 228C
51.6 206D
61.5 m246
65.4 m491
69.9 m112
80.5 m498
LG20
0.0 227A
9.2 NYU15
21.8 208A
31.4 m360
33.9 m361
42.1 113C
68.1 m420
74.8 m359
LensE
LEN
COMD
LG21
0.0 216E
3.6 m323
6.9 m324
34.6 m318
46.6 m512
49.5 m325
50.1 m320
56.0 [m319] m326
66.2 m537
77.5 m321
78.9 m125
83.3 m483
97.5 m120
LG22
0.0 122B
6.7 207B
29.3 m429
34.5 m419
36.4 m266
WtLoss


LG23
0.0 202D
6.5 213B
13.8 24B
LG24
0.0 237E
4.6 218B
13.8 m34
LG25
0.0 2B
8.2 NYU32
20.1 m252
21.0 m257
28.6 m476
37.8 m10
39.6 m251
40.9 m254
43.0 m256
44.8 m470
48.3 m39
49.9 m116
54.0 m452
57.6 m253
62.7 216C
74.2 m432
80.2 m431
83.2 m255
RelDent
MEL_E
LG26
0.0 136B
6.5 NYU33
10.4 m200
12.8 m195
15.4 m194
16.2 m12
17.4 m196
20.2 m192
24.5 m508
26.3 m16
30.0 Six3
COND
AASense
EYE_M
LG27
0.0 m150
1.6 m467
3.5 m265
18.0 m121Pl
EYE
LG28
0.0 BMP4
0.8 m224
4.6 206A
11.9 m219
16.2 m221
21.8 209E
26.5 m222
36.2 POMCsnp1
45.1 Otx1
51.5 m223
60.4 m220
WtLoss
COND
LG29
0.0 229B
9.1 223C
19.2 235D
29.6 m243
EYE
COND
LG31
0.0 m259
5.7 m261
11.1 6C
15.5 214D
29.1 226E
39.5 214C
48.1 m260
56.0 m94
60.1 m258
88.5 m80
95.5 m505
105.6 215A
LenL
COND
MEL_D
LG32
0.0 m212
9.4 m216
LG33
0.0 m66
2.8 m443
4.0 m487
[m232] [m235]
5.8 m230
7.2 m233
16.9 m231
WtLoss
COND
LG34
0.0 131C
6.8 m479
13.0 m237
14.7 m239
15.3 m500
20.4 m238
25.1 m503
LEN
MEL_D
MEL_A
LG35
0.0 55B
6.4 m397

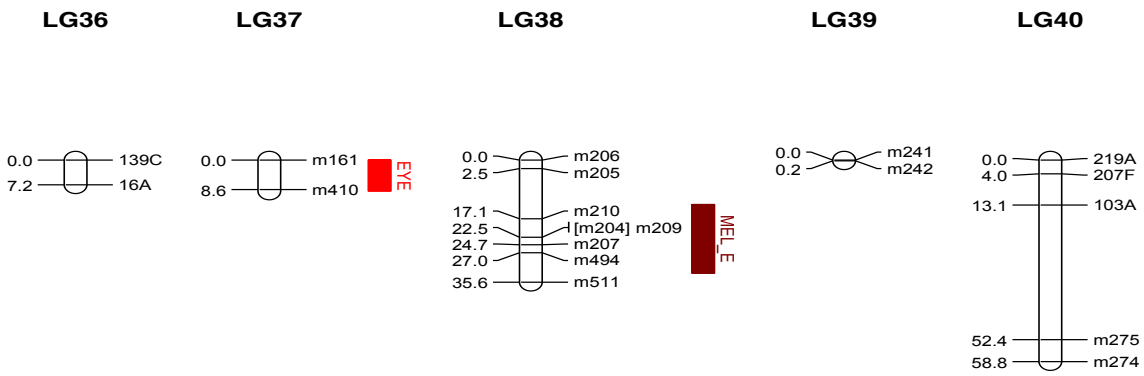



Figure 3.2. Integrated linkage maps (microsatellite + SNPs) of *Astyanax mexicanus* with colored bars denoting positions of detected QTL for specific trait. Marker positions are given in cM; QTL bar denotes one LOD score confidence intervals.

| Trait (N) | Symbol | Description and trait |
|---|---|---|
| Eye size | RelEye | Observed eye size divided by eye size predicted from the regression of eye size on standard length (SL). Skew corrected in MultiQTL. Cave < surface. Wilkens (1988) |
| Pigmentation | EyesMel | Count of melanophores in an area (2.0×0.4 mm) located 3.0 mm above the left eye. See supplemental methods in Protas et al. 2007 for illustration. Cave < surface. Wilkens (1988) |
| Pigmentation | DorsalMel | Count of dorsal melanophores in an area (2.0×0.4 mm) located 3.0 mm above the left eye. See supplemental methods in Protas et al. 2007 for illustration. Cave < surface. Wilkens (1988) |
| Lens size | LenL | Observed lens size divided by lens size predicted from the regression of lens size on standard length (SL). Skew corrected in MultiQTL. Cave < surface |
| Relative condition | COND | Observed weight divided by predicted weight calculated from regression of log weight on log length. Skew corrected in MultiQTL. Surface < cave |
| Weight loss | Wtloss | Rate of weight loss on fast expressed as percent decrease per day. Weight loss is slower in cave than in surface individuals. Huppop (1986) |
| Ribs | ribs | The number of thoracic ribs. Surface fish have 12, Pachón cave fish have 11 or 12. Dowling et al. (2002) |
| Length | GrLEN | Estimated daily growth rate assuming that all the fry started at the same length (starting at 4 mm). Because the F2 was comprised of individuals from different broods, the use of specific time periods for measurements, which differed among the broods,we |
| Length | ResidLen | ResidLen was the residual of a ANOVA in which length was the dependent variable and individual qualitative group identifiers were the independent variables. Thus, the residual has group differences removed. |
| Chemical sense | AAsense | Threshold sensitivity to dissolved amino acids in the water assessed by searching response triggered by the addition of amino acid solution to the test aquarium. Responses were scored over the full range of concentrations and used to estimate the concen |

Table 3.1.A. Summary and description of the measured phenotypes, abbreviations and their mean values in $F_2$ generation as measured by Protas [77].

| Trait | Ch | LOD | P- value | Pos (cM) | PEV | PEVad |
|---|---|---|---|---|---|---|
| AASens | LG11 | 5.08 | 0.001 | 0 | 0.102 | 0.053 |
| AASens | LG15 | 5.52 | 0.001 | 0 | 0.089 | 0.001 |
| AASens | LG16 | 3.59 | 0.002 | 12.8 | 0.068 | 0.004 |
| AASens | LG17 | 4.54 | 0.002 | 59.2 | 0.084 | 0.079 |
| AASens | LG26 | 8.36 | 0.001 | 20.2 | 0.271 | 0.117 |
| AASens | LG4 | 4.83 | 0.001 | 85.3 | 0.082 | 0.003 |
| DorsalMel | LG1' | 4.89 | 0.001 | 74.1 | 0.105 | 0.019 |
| DorsalMel | LG12 | 3.94 | 0.002 | 38.9 | 0.238 | 0.166 |
| DorsalMel | LG13 | 15.5 | 0.001 | 27.4 | 0.272 | 0.271 |
| DorsalMel | LG31 | 6.88 | 0.001 | 35.3 | 0.067 | 0.048 |
| DorsalMel | LG34 | 2.95 | 0.006 | 15.3 | 0.032 | 0.014 |
| EyesMel | LG1' | 3.97 | 0.001 | 78.2 | 0.071 | 0.036 |
| EyesMel | LG13 | 15.1 | 0.001 | 30.4 | 0.348 | 0.348 |
| EyesMel | LG25 | 4.14 | 0.001 | 42.4 | 0.088 | 0.088 |
| EyesMel | LG26 | 4.1 | 0.002 | 30.0 | 0.071 | 0.050 |
| EyesMel | LG38 | 3.88 | 0.001 | 22.6 | 0.050 | 0.028 |
| GrLen | LG34 | 3.27 | 0.004 | 7.1 | 0.071 | 0.045 |
| RelCond | LG20 | 5.14 | 0.005 | 61.9 | 0.062 | 0.046 |
| RelCond | LG26 | 3.83 | 0.005 | 30.0 | 0.041 | 0.041 |
| RelCond | LG28 | 5.0 | 0.005 | 2.8 | 0.064 | 0.060 |
| RelCond | LG29 | 5.24 | 0.005 | 25.7 | 0.089 | 0.021 |
| RelCond | LG31 | 5.78 | 0.005 | 15.5 | 0.051 | 0.044 |
| RelCond | LG33 | 4.64 | 0.005 | 9.5 | 0.051 | 0.047 |
| RelCond | LG8 | 4.26 | 0.010 | 11.5 | 0.056 | 0.039 |
| RelEye | LG1' | 13.5 | 0.001 | 132.1 | 0.074 | 0.050 |
| RelEye | LG12 | 4.78 | 0.001 | 21.5 | 0.023 | 0.017 |
| RelEye | LG14 | 56.98 | 0.001 | 35.0 | 0.481 | 0.474 |
| RelEye | LG1 | 5.95 | 0.001 | 88.7 | 0.033 | 0.029 |
| RelEye | LG27 | 3.68 | 0.001 | 0 | 0.016 | 0.000 |
| RelEye | LG29 | 9.38 | 0.001 | 2.7 | 0.056 | 0.056 |
| RelEye | LG37 | 10.0 | 0.001 | 0.2 | 0.025 | 0.025 |
| RelEye | LG3 | 3.5 | 0.002 | 11.6 | 0.015 | 0.003 |
| ResidLen | LG10 | 4.24 | 0.004 | 13.3 | 0.058 | 0.052 |
| ResidLen | LG14 | 2.99 | 0.004 | 18.6 | 0.047 | 0.045 |
| ResidLen | LG20 | 3.1 | 0.004 | 14.3 | 0.074 | 0.027 |
| ResidLen | LG35 | 10.0 | 0.004 | 6.4 | 0.071 | 0.063 |
| ResidLen | LG4 | 2.59 | 0.008 | 0 | 0.034 | 0.000 |
| ribs | LG31 | 7.27 | 0.001 | 19.9 | 0.117 | 0.116 |
| ribs | LG37 | 10.0 | 0.001 | 8.3 | 0.051 | 0.004 |
| ribs | LG3 | 5.67 | 0.001 | 0.10 | 0.103 | 0.027 |
| ribs | LG40 | 13.34 | 0.001 | 58.8 | 0.301 | 0.295 |
| ribs | LG4 | 4.33 | 0.001 | 11.6 | 0.054 | 0.047 |
| WtLoss | LG22 | 3.56 | 0.001 | 4.4 | 0.089 | 0.049 |
| WtLoss | LG28 | 3.54 | 0.001 | 21.8 | 0.056 | 0.040 |
| WtLoss | LG33 | 2.91 | 0.003 | 2.8 | 0.060 | 0.034 |
| WtLoss | LG3 | 4.38 | 0.003 | 198.8 | 0.082 | 0.016 |
| LenL | LG20 | 4.287 | 0.001 | 35.6 | 0.071 | 0.014 |
| LenL | LG3 | 8.815 | 0.001 | 103.1 | 0.134 | 0.057 |
| LenL | LG14 | 11.05 | 0.001 | 65.3 | 0.169 | 0.158 |

Table 3.1.B. Summary of identified QTL with their respective linkage groups (Ch), position in centimorgans (Pos), maximum LOD score (LOD) and P-values (P-value). Permutation was used to asses significances of all QTL and confidence intervals on their positions were determined by bootstrap analyses. Significance threshold was set at P = 0.05 for individual QTL, with a genome-wide false detection rate of 10% (FDR = 0.10) PEV and PEVad refer to the proportions of phenotypic trait variance in the mapping progeny ($F_2$) that

are explained by a QTL. PEV refers to total trait variance; PEVad refers to the proportion of additive variance explained by the QTL.

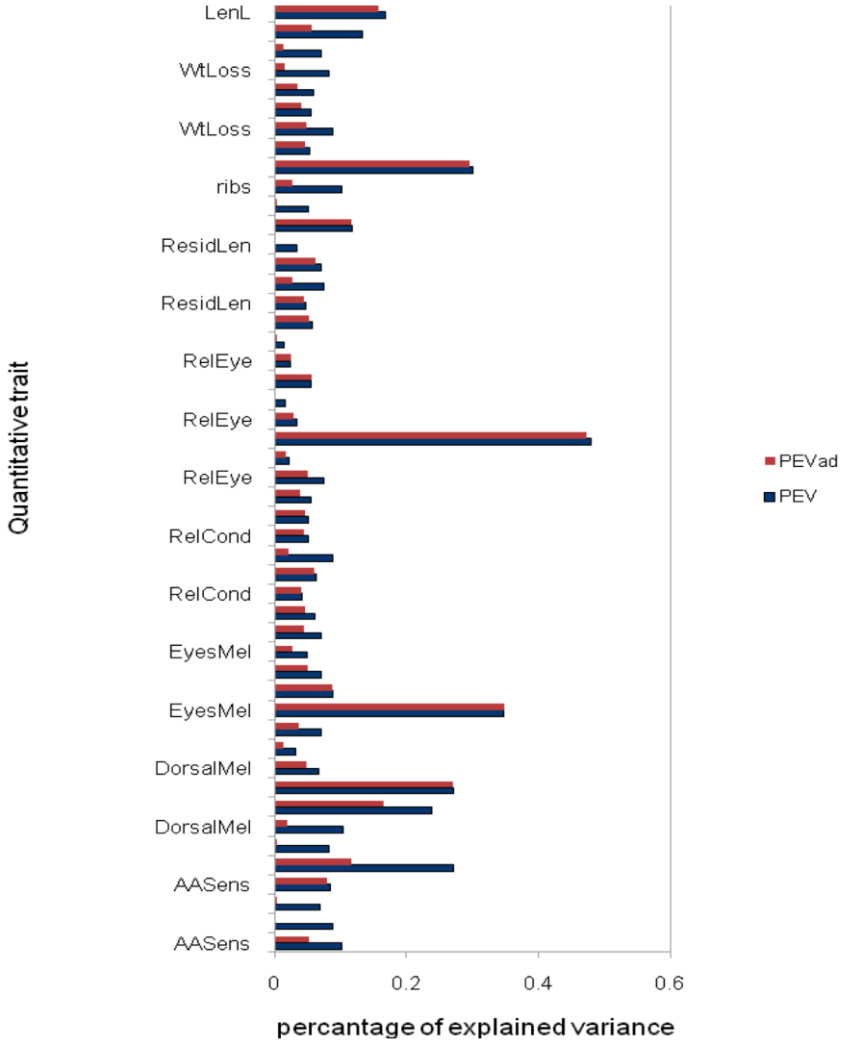



Figure 3.3. Distribution of the percentages of total (PEV) and additive (PEVad); variance explained (PVE) at the phenotypic loci per each trait on the QTL map as identified using MultiQTL (see methods for details). X-axis represents percentage of either PEV or PEVad, while Y-axis represents different QTL for studied traits. Color of the bars is identifying PEV (blue) or PEVad (orange). It must be noted that one trait might have several QTLs, which are represented separately on the graph.

**Diversity in natural populations and ascertainment bias**

*Quality control of surveyed markers*

To perform a genome wide screen of the natural populations of *A. Mexicanus* we collected 281 individuals from 12 populations and scored them for 745 SNP markers. Samples were pooled as described in material and methods. First, we applied quality control to the entire data leading to our discarding 80% of the missing markers and individuals as well as all the monomorphic loci. This analysis retained 272 individuals and 519 high quality genotypes which was ~70% of the total genotyped markers. We did not observe any difference in the success of genotyping assays while comparing Sanger sequencing derived markers with RAD tag sequencing technology (data not shown). Table 3.2 lists the loci scored for each population after first level of quality control. The highest number of the polymorphic loci were present in surface populations (SN1, SN2) as well as in the admixed cave populations N3* and O8*; these numbers were 326, 381, 341, 396, respectively (*note that asterisk next to the population name signifies admixed population and this notation will be retained through the chapter. Some of the populations are combined as described in Materials and Methods at the end of the chapter*).

We performed additional quality controls, discarding markers for which the less common alleles had frequencies lower than 5% per population (minor allele frequency, MAF < 5%) to evaluate the presence of the common polymorphic loci (MAF > 5%). This procedure reduced the number of polymorphic markers under consideration due to the low frequencies (MAF < 5%) of the polymorphism in the cave populations (N1 = 43, N2 = 154, O4O6 = 143). The surface population that was previously hypothesized to be the closest to the old cave populations exhibited an even lower number of polymorphic markers than most of the cave populations (SO = 76). This was also due to the presence of many low polymorphic loci (MAF < 5%). The distributions of MAF per each population are showed in the supplementary material (Figure S3.2).

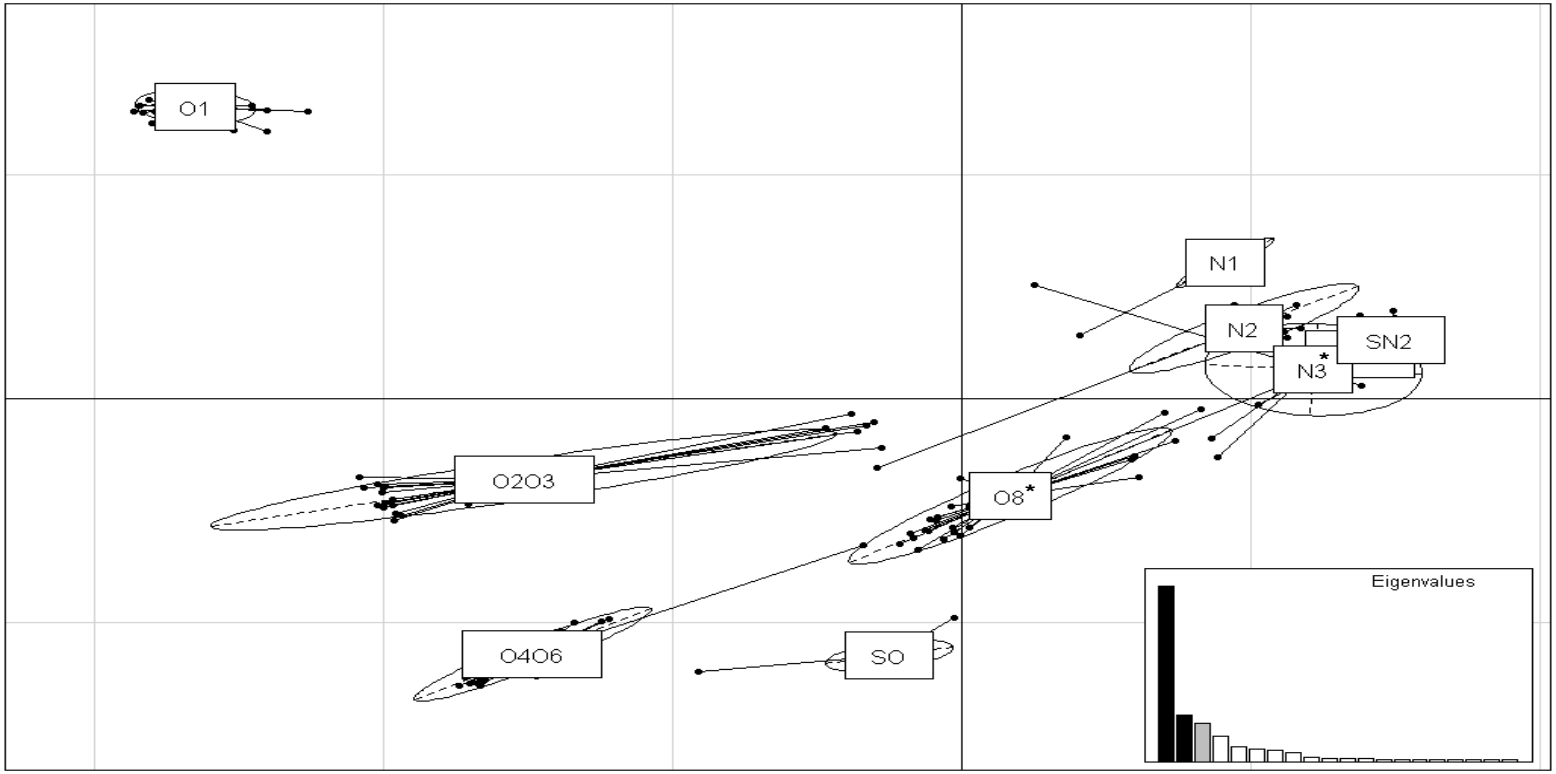
We also performed Hardy-Weinberg equilibrium (HWE) tests on each marker deviations from HWE that is discussed later in the chapter.

| Population | Abbreviation | Origin | N | a | b |
|---|---|---|---|---|---|
| El Abra and Guatemala surface localities SN1 (S1, S2, S3) | **SN1** | surface | 25 | 326 | 255 |
| Rio Subterráneo Valley | **SN2** | surface | 45 | 381 | 292 |
| Cueva Molino | **N1** | New cave | 22 | 131 | 43 |
| Cueva Caballo Moro | **N2** | New cave | 26 | 270 | 154 |
| Cueva Micos | **N3*** | New cave | 25 | 341 | 242 |
| Cueva Pachón | **O1** | Old cave | 31 | 345 | 186 |
| Cueva Jerbaniz and Japonés | **O2O3** | Old cave | 22 | 375 | 291 |
| Cueva Arroyo and Curva | **O4O6** | Old cave | 29 | 277 | 143 |
| Cueva Chica | **O8*** | Old cave | 32 | 396 | 324 |
| Rascon | **SO** | surface | 24 | 178 | 76 |

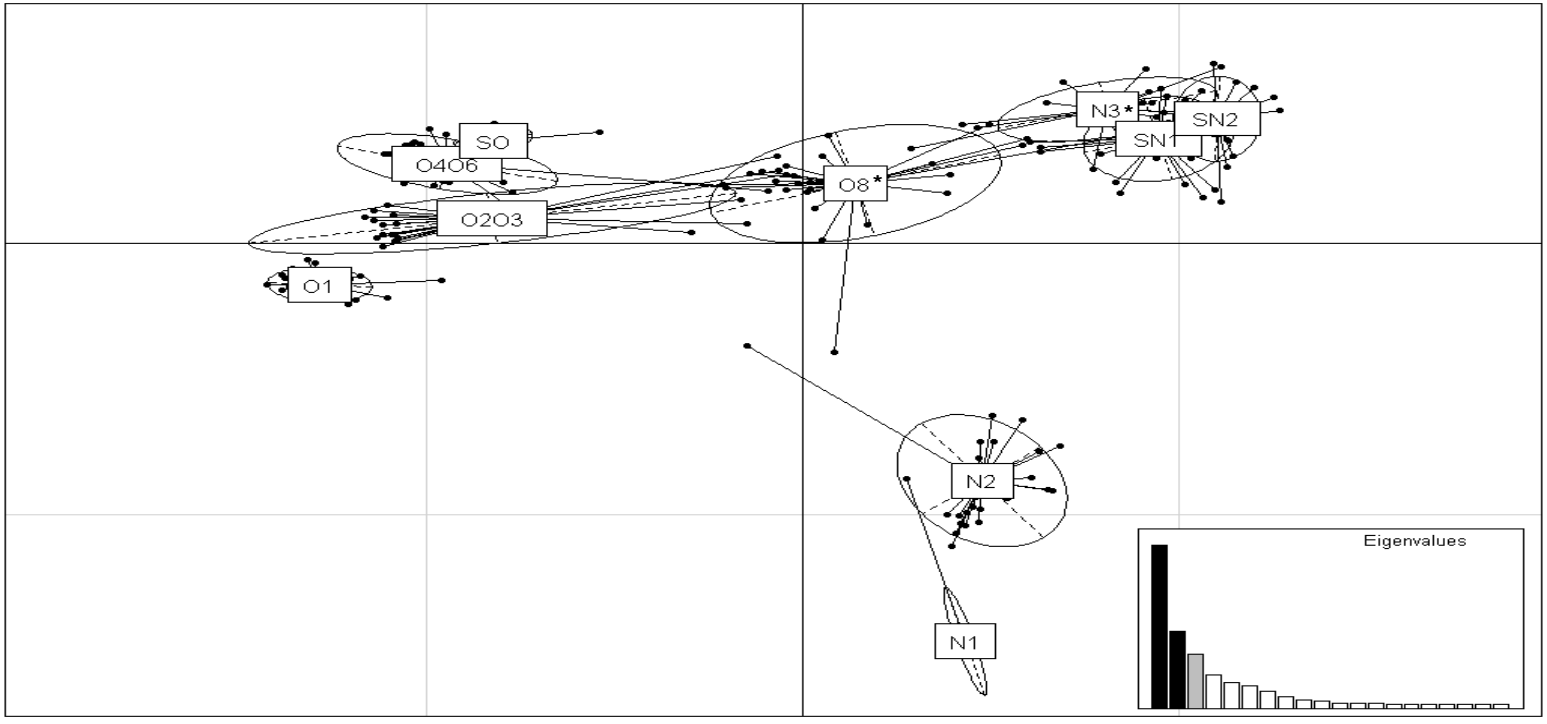
Table 3.2. Details on sample locations, population abbreviations, origin, sampled individuals (N), and marker quality control per populations (a and b) (See methods for details). We show here two steps of the quality control per each population after the quality control applied on the entire data set that reviled 519 markers (markers that were monomoprhic or did not amplify in all the populations were removed): a. the number of the SNPs excluding those that were monomorphic in respective population, b- the number of the SNPs excluding those that were MAF<5% in respective population. Each of the quality control was done per population (See Materials and methods for details). Asterisk (*) determines admixed populations previously described in Chapter 2.

*Ascertainment Bias estimation in different SNP panels*

In order to determine if discovered set of SNPs was biased and if this could affect intrinsic differences in diversity levels (ascertainment bias) we explored population genetic parameters for the groups of markers across the populations. The full set of SNPs was used in order to evaluate structure of the studied populations using Principal Components Analysis (PCA). The PCA analysis revealed major divergence between the new and old lineages as well as significant regional structuring of the old populations. The PCA 1 and 2 axes explained 51.56% and 13.74% of the variation, respectively (Figure 3.4.A). We noticed the unusual grouping of the O1 population, which provided a first insight of the potential biases that could have been produced by SNP discovery. The O1 population is of old origin and an individual from it was a parent in the F1 cross from which the RAD-tag SNPs were developed. Thus, in order to understand this source of ascertainment bias, the RAD-tag SNP loci with the MAF<5% in the surface population were removed and PCA was performed again. This correction revealed again major population structuring between the old and new lineages and a smaller, but still significant, differentiation among the old populations. An, additional differentiation emerged between new populations (N1, N2) represented in two PCA axis explaining 18.91% and 8.90%, respectively (Figure 3.4.B).




3.4.A. PCA for all the SNP markers used in the study. Values in parentheses indicate variance explained by two coordinate eigenvalues (51.6 %, 13.7 %, and 11.2 %, respectively). Populations are identified by their abbreviation; old caves (O1, O2O3, O4O6, O8*), new caves (N1, N2, N3*), surface population (SN1, SN2) and old surface population (SO).




3.4.B. PCA for SNPs produced by RAD tag and Sanger sequencing method for the markers with MAF>5% in surface populations. PCA eigenvalues explain: 18.9, 8.9, and 6.3%, respectively. Populations are identified by their abbreviation; old caves (O1, O2O3, O4O6, O8*), new caves (N1, N2, N3*), surface population (SN1, SN2,) and old surface population (SO).

Based on the above mentioned pattern all the markers were divided into the groups 1) SNP (high polymorphic, MAF > 5%) and 2) low polymorphic (MAF<5%) in the surface population, and the expected heterozygosities ($H_e$) for each population and each group of markers (Figure 3.5) were calculated. These two groups were each further divided in two additional groups based on the sequencing method (RAD tag or Sanger sequencing) in order to observe the potential differences. The groupings of the SNPs (high polymorphic (MAF>5%) was based on the hypothesis that the majority of the variability should be present in the surface population in the form of standing genetic variation. Surface populations (SN1, SN2) besides admixed cave populations (N3* and O8*) had the highest number of the polymorphic loci after removing MAF < 5% (Table 3.3). Descriptive statistics for each marker group and population are available in Table 3.3.

Trends in heterozygosity differ between two groups of markers conditioned on whether the polymorphism was found in surface populations (surface RAD and surf seq MAF > 5%) or was very low in the surface populations (surface RAD and surf seq MAF < 5%). Thus, markers were further collapsed in two main groups and we will refer to them throughout the chapter as “surface SNPs” which represents RAD tag and sequenced derived markers with MAF>5% in the surface populations and “cave SNPs” which represent RAD tag and sequenced markers represented by low variants (MAF < 5%) in the surface populations.

It is important to notice that the heterozygosity in some populations was very low regardless of the marker groups (N1, O4O6, SO). Also the old cave populations show very similar polymorphism level when compared to new cave populations when only “surface SNPs” were considered. The surface fish population (SO) showed very low polymorphism in both groups of markers and could not be placed properly in order to test our hypothesis, thus SO was not considered in further analyses. Based on these observations we have

assumed in our further analysis that the variation in “surface SNPs” (from SN1 and SN2 populations) is a good representative of the ancestral polymorphism in new cave lineages and is shared to some extend with the old cave lineage. “cave SNPs” are present in most of the old cave populations and are specific for the old lineage.

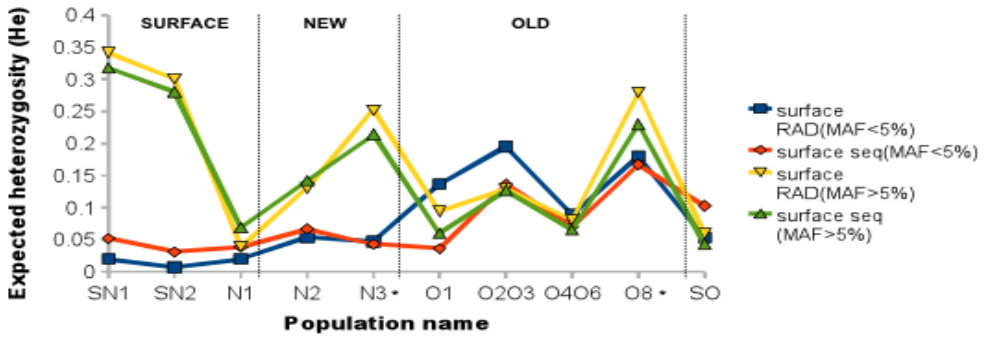



Figure 3.5. Trends in heterozygosity for different marker panels. Dashed line separates different groups of the population origins with the population names designated on the x-axis. Different line colors represent different marker groups. Markers are divided in four groups; two groups based on the sequencing methodology used (RAD and Sanger sequencing) which are further divided in two more groups based on MAF in surface populations (MAF > 5% or MAF < 5%), as described in the legend. Note that asterisk (*) next to the population name signifies admixed population.

*Genetic Diversity in the populations*

We calculated descriptive statistics by averaging population genetic parameters across all the markers. Mean observed heterozygosity ($H_o$) averaged over both sets of markers ranged from 0.05 ± 0.11 (N1) to 0.19 ± 0.11 (N3*) and the mean unbiased expected heterozygosity (Nei's $H_e$) varied from 0.05 ± 0.09 (N1) to 0.19 ± 0.12 (N3*) (Table 3.3). Mean allelic richness (AR) varied from 1.1 ± 0.08 (N1) to 1.47 ± 0.41 (O8*). As expected, the proportion of polymorphic loci (%P) was positively correlated with the observed heterozygosity (Spearman's coefficient: S = 24.0, p-value = 0.0096). The average proportion of polymorphic loci was highly variable among populations

(%P) and ranged from 13.45 ± 0.015 (O4O6) to 49.66% ± 0.47 (N3*). Overall we detected the greatest diversities based on heteroyzgosity ($H_e$ and $H_o$) and allelic richness in the surface and admixed cave populations (N3* and O8*) (Table 3.3). Private allelic richness (the alleles that are found in only a single population) was detected in surface populations (~0.01) and in the old cave population (ranged from 0 for O8* to 0.10 for O1) while there was no private allele observed in the new cave populations (Table 3.3). We did not detect significant difference of private alleles content when comparing surface and old caves (Wilcoxon test, W = 4, p-value = 0.6831). However, it should be noted that population of the old origin (O1) is characterized by the highest number of private alleles (10%, Table 3.3). This suggests that there are markers that are highly specific for the old cave populations (i.e. O1) or, alternatively, are present in the surface population but in such low frequency as to elude detection. Averaged inbreeding coefficients across populations ($F_{IS}$) ranged from -0.02 ± 0.04 in O1 population to 0.31 ± 0.33 in the new cave populations (N2 and N3*) (Table 3.3).

| Population | MA nb. | AR | PAR | $H_o$ | Nei's $H_e$ | % P | $F_{IS}$ |
|---|---|---|---|---|---|---|---|
| "Surface markers" | | | | | | | |
| SN1 | 1.927 | 1.89 | 0.01 | 0.237 | 0.267 | 92.7393 | 0.114305 |
| SN2 | 1.878 | 1.78 | 0.01 | 0.281 | 0.295 | 87.7888 | 0.048179 |
| N1 | 1.564 | 1.16 | 0 | 0.081 | 0.085 | 56.4356 | 0.045029 |
| N2 | 1.851 | 1.45 | 0 | 0.214 | 0.242 | 85.1485 | 0.116468 |
| N3* | 1.99 | 1.71 | 0 | 0.31 | 0.335 | 99.0099 | 0.077509 |
| O1 | 1.706 | 1.34 | 0 | 0.134 | 0.128 | 70.6271 | -0.04225 |
| O2O3 | 1.32 | 1.5 | 0 | 0.052 | 0.055 | 32.0132 | 0.061946 |
| O4O6 | 1.264 | 1.3 | 0 | 0.052 | 0.047 | 26.4026 | -0.115107 |
| O8* | 1.614 | 1.76 | 0 | 0.11 | 0.133 | 61.3861 | 0.179677 |
| "Cave markers" | | | | | | | |
| SN1 | 1.301 | 1.01 | 0 | 0.019 | 0.024 | 0.300926 | 0.203388 |
| SN2 | 1.06 | 1.02 | 0 | 0.008 | 0.01 | 0.060185 | 0.176811 |
| N1 | 1.199 | 1.04 | 0 | 0.014 | 0.022 | 0.204651 | 0.372517 |
| N2 | 1.356 | 1.06 | 0 | 0.028 | 0.055 | 0.356481 | 0.496121 |
| N3* | 1.319 | 1.05 | 0 | 0.022 | 0.047 | 0.319444 | 0.539675 |
| O1 | 1.764 | 1.12 | 0.19 | 0.124 | 0.125 | 0.763889 | 0.005972 |
| O2O3 | 1.708 | 1.19 | 0.02 | 0.192 | 0.188 | 0.708333 | -0.021251 |
| O4O6 | 1.505 | 1.09 | 0.01 | 0.061 | 0.087 | 0.50463 | 0.299363 |
| O8* | 1.495 | 1.18 | 0 | 0.14 | 0.178 | 0.49537 | 0.219161 |
| **AVERAGE OVER "surface and "cave" markers** | | | | | | | |
| **SN1** | 1.614 | 1.45 | 0.005 | 0.128 | 0.1455 | 46.520113 | 0.1588465 |
| **SN2** | 1.469 | 1.4 | 0.005 | 0.1445 | 0.1525 | 43.9244925 | 0.112495 |
| **N1** | 1.3815 | 1.1 | 0 | 0.0475 | 0.0535 | 28.3201255 | 0.208773 |
| **N2** | 1.6035 | 1.255 | 0 | 0.121 | 0.1485 | 42.7524905 | 0.3062945 |
| **N3*** | 1.6545 | 1.38 | 0 | 0.166 | 0.191 | 49.664672 | 0.308592 |
| **O1** | 1.735 | 1.23 | 0.095 | 0.129 | 0.1265 | 35.6954945 | -0.018139 |
| **O2O3** | 1.514 | 1.345 | 0.01 | 0.122 | 0.1215 | 16.3607665 | 0.0203475 |
| **O4O6** | 1.3845 | 1.195 | 0.005 | 0.0565 | 0.067 | 13.453615 | 0.092128 |
| **O8*** | 1.5545 | 1.47 | 0 | 0.125 | 0.1555 | 30.940735 | 0.199419 |

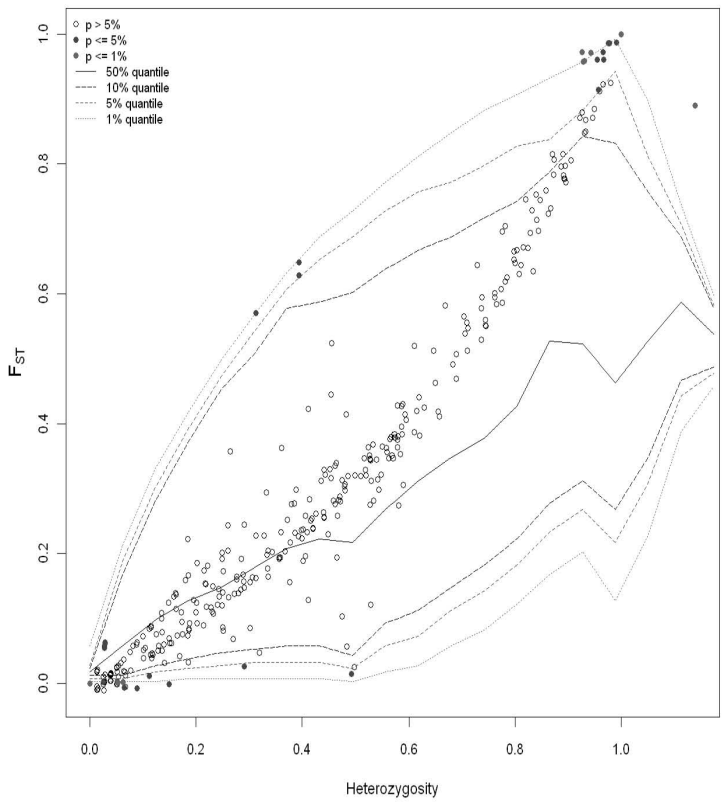
Table 3.3. Genetic diversity for two marker groups ("cave SNPs" and "surface SNPs" markers) and averaged parameters per populations. For each population summary statistic measures are given for each marker panel and the average over entire marker set. The abbreviations for the measures are following: N-number of individuals, MA nb- mean allele number, AR-allelic richness, PAR-private allelic richness; $H_o$-observed heterozygosity, Nei $H_e$-expected heterozygosity standardized by sample size per each population according to Nei 1978. % P-percentage of the polymorphic loci obtained per each population for each group of markers, $\mathbf{F_{IS}}$ -deviations from random mating. Note that asterisk next to the population name signifies admixed population.

**Population genetic parameters estimates on single loci**

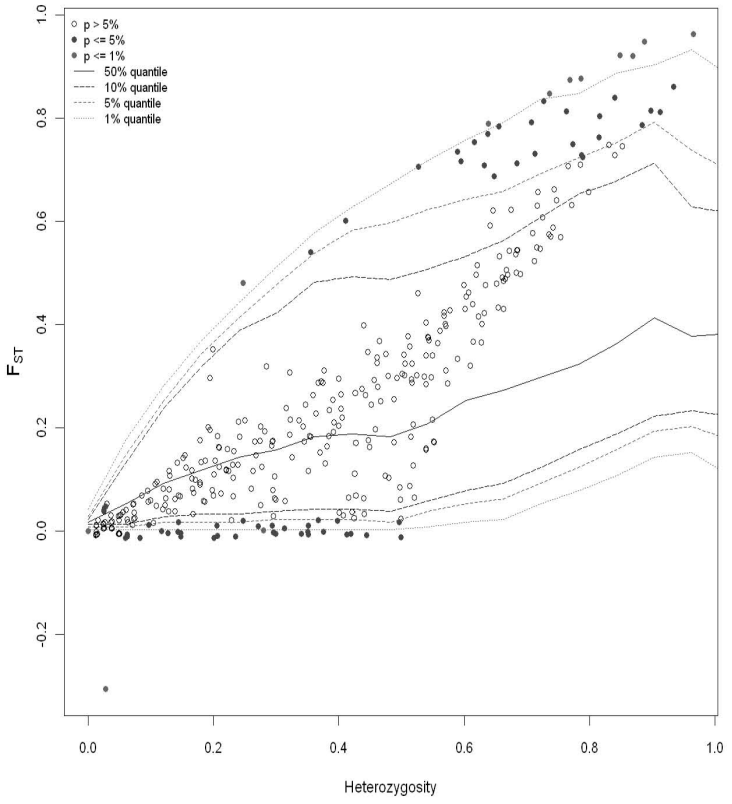
*Differentiation among populations and outlier locus detection*

Detection of outlier SNP loci and the distribution of empirical $F_{ST}$ values were performed for all identified SNPs described above according to the hierarchical model implemented in ARELQUIN 3.5 [213]. We used a hierarchical island model because it is not influenced by population structure [165]. The hierarchical island model also uses nested variance analysis and allows for higher exchange of migrants between the demes within groups than between groups. Thus, it is superior method to use in this aspect of our study because we have already resolved the population structure and showed that migration exchanges are unequal among populations (Chapter 2). The null distribution of the tested loci generated under the hierarchical island model is summarized by quantiles of the joint distribution for each population pair in Figure 3.6. Locus specific $F_{ST}$ and p-values are also shown on figure 3.6 indicating 1% and 5% outliers represented as red and blue filled circles, respectively and summarized in the supplementary material (Table S3.2, Supplementary material).

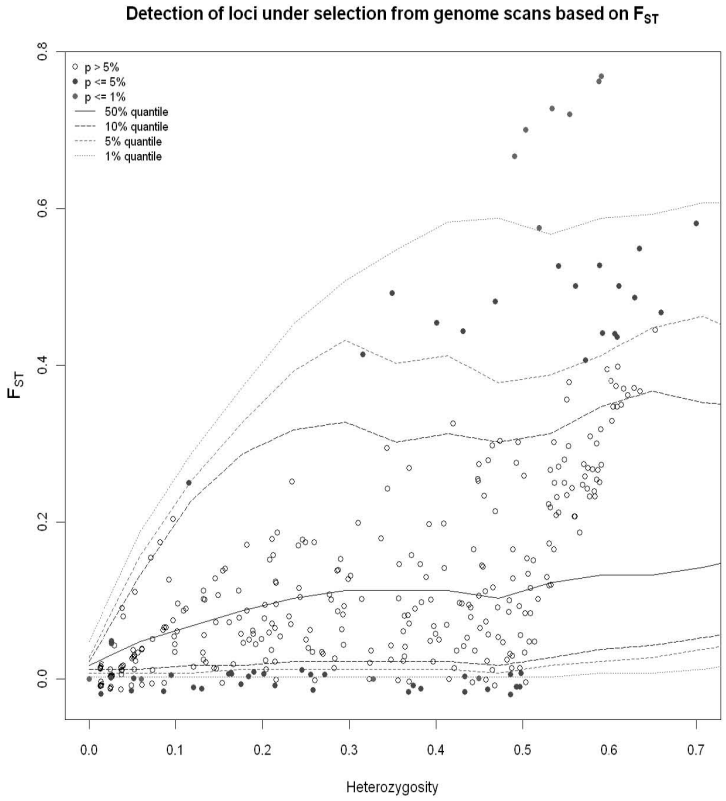
A


Detection of loci under selection from genome scans based on $F_{ST}$
p > 5%
p <= 5%
p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
$F_{ST}$
Heterozygosity
0.0
0.2
0.4
0.6
0.8
1.0

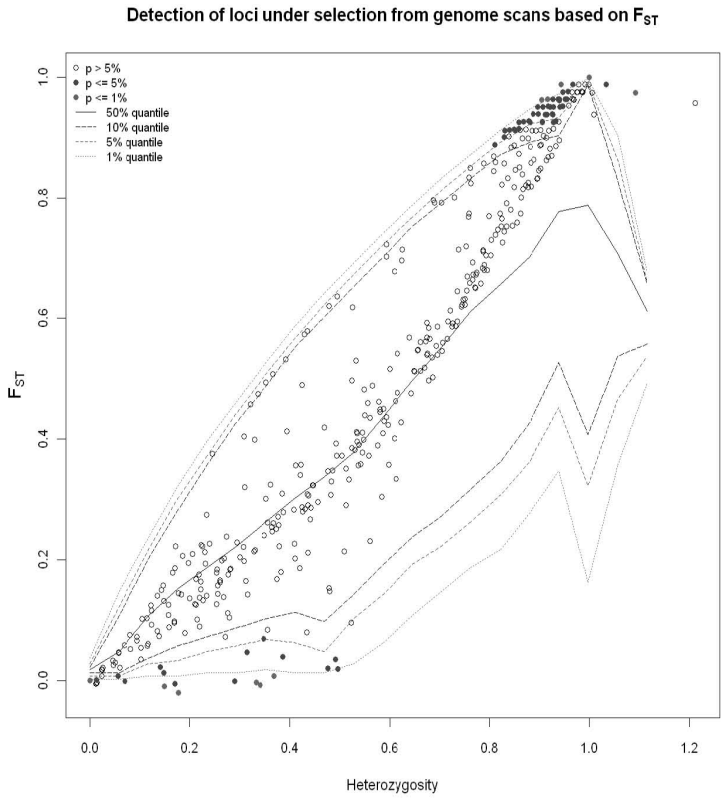
B.


Detection of loci under selection from genome scans based on $F_{ST}$
p > 5%
p <= 5%
p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
$F_{ST}$
Heterozygosity
-0.2
0.0
0.2
0.4
0.6
0.8
1.0

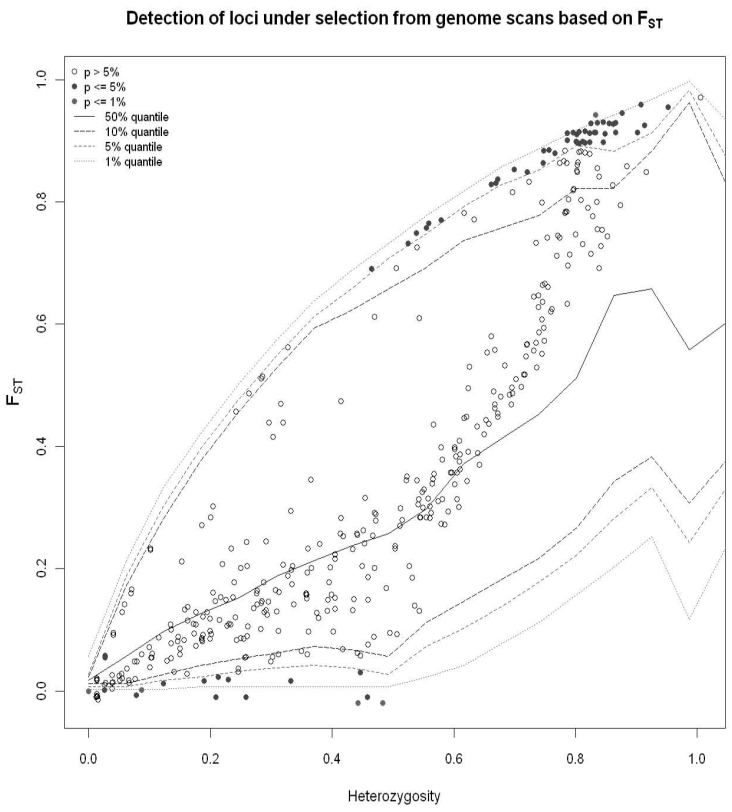
C.


Detection of loci under selection from genome scans based on $F_{ST}$
p > 5%
p <= 5%
p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
$F_{ST}$
Heterozygosity
0.0
0.2
0.4
0.6
0.8
0.1
0.3
0.5
0.7

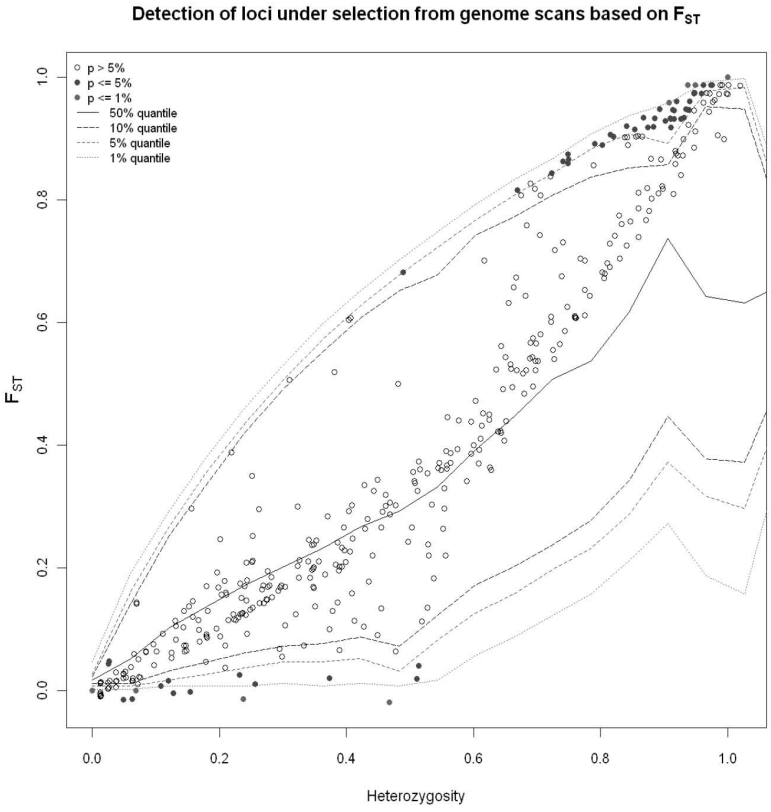
D.


Detection of loci under selection from genome scans based on $F_{ST}$
p > 5%
p <= 5%
p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
$F_{ST}$
Heterozygosity
0.0
0.2
0.4
0.6
0.8
1.0
1.2

E


Detection of loci under selection from genome scans based on $F_{ST}$
p > 5%
p <= 5%
p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
$F_{ST}$
Heterozygosity
0.0
0.2
0.4
0.6
0.8
1.0

F.


Detection of loci under selection from genome scans based on $F_{ST}$
p > 5%
p <= 5%
p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
$F_{ST}$
Heterozygosity
0.0
0.2
0.4
0.6
0.8
1.0

G.

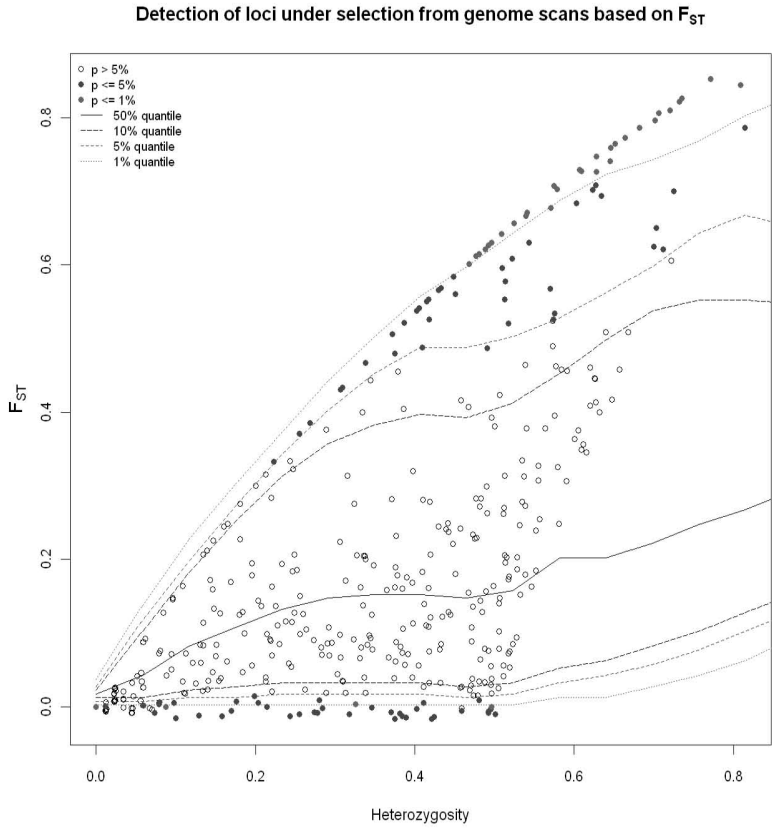

Detection of loci under selection from genome scans based on $F_{ST}$
p > 5%
p <= 5%
p <= 1%
50% quantile
10% quantile
5% quantile
1% quantile
$F_{ST}$
0.0
0.2
0.4
0.6
0.8
Heterozygosity

Figure 3.6. Distributions of the outlier loci in multiple cave *vs.* surface comparisons. Per locus $F_{ST}$ values were calculated at individual markers, with the following population structure: Surface (SN1, SN2) first group, respective cave second group. $F_{ST}$ is represented as a function of expected heterozygosity for all the population pairs. Hierarchical island model is used to produce the null distribution. The solid line represents the 50% quantile while the blue dashed line represents the 5% and the 95% quntiles of the distribution. Red dotted line represents 1% and 99% of the distribution. Each empty circle represents a single SNP marker. Red and blue filled circles represent differentiated markers with 1% or 5% significance, respectively A. N1 *vs.* surface (SN1, SN2), B. N2 *vs* surface (SN1, SN2). C. N3* *vs.* surface (SN1,SN2), D. O1 *vs.* surface (SN1,SN2), E. O2O3 *vs.* surface (SN1,SN2), F. O4O6 *vs.* surface (SN1,SN2), G. O8* *vs.* surface (SN1,SN2). SNPs that had $F_{ST}$ values above the 95% and below 5% of the confidence intervals were considered as outlier loci.

The analysis identified numerous loci, depending on surface-cave comparison, as strongly differentiated. We have identified loci under putative balancing and directional selection. Although the concept of balancing selection is well established [158] there are still methodological limitations for the identification of balancing selection [165, 168, 230]. In our analysis, majority of the loci that were identified as putative balancing selection were present in the admixed populations and were not consistent between the populations (N3* and O8*). Therefore, the following treatment focuses only on the loci putatively under directional selection. Relative to the surface populations, the percentages of outlier loci ranged from ~ 6% for N1 population to as high as 25% of the total scanned loci in O1 population. In general, the number of outlier loci detected from an individual SNP marker was higher in old cave populations and significantly different between two groups of populations (new *vs.* old $\chi^2$ = 72.653, DF = 6, p-value = 1.2e-13) (Table 3.3). However, a significant portion of the loci identified in those populations came from the set of markers that have been identified as a low polymorphism in surface populations (SN1, SN2) (MAF<5%) ($\chi^2$= 242.7, DF = 7, p-value < 2.2e-16). The same was observed for markers from the new cave populations ($\chi^2$ = 75.0, DF = 5, p-value = 9.1e-15). Significance was observed for all individual comparisons of each cave populations within the new origin and among new and old origin (not shown).

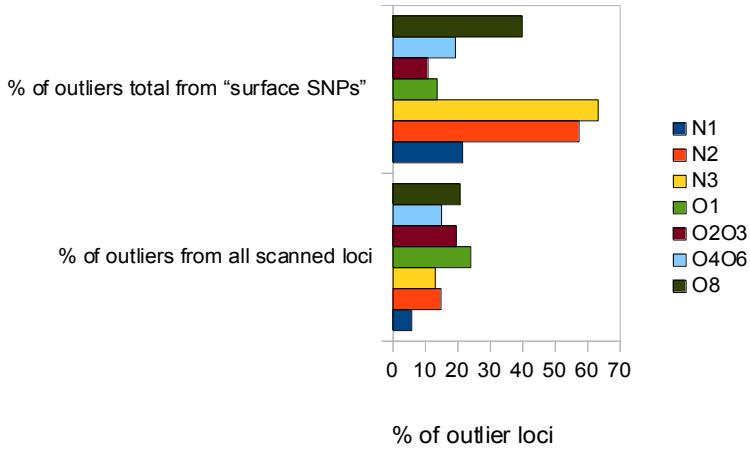



Figure 3.7. Summary of the differentiated loci detected by hierarchical outlier test per each population. Loci are represented as the percentage (%) of the detected outliers under directional selection from all the loci included in the study % of outliers coming only from “surface SNPs”. Different populations are color-coded.

There was little evidence for strong differentiation of the candidate loci in multiple cave populations using $F_{ST}$ based outlier test. For example we have found differentiation of only one (m682) out of five SNPs from the Oca2 gene just in old cave populations (O1, O2O3, O4O6, O8*). Another example is that of two SNPs from the BACGH that contained the potential candidate, growth hormone gene. Diversification of these markers (m595, m619) was also observed only in old cave populations (N2, N3*) (Table 3.4). We did not find differentiation of any SNP markers in the αA-crystallin gene, Mc1r or Prox1 gene that were chosen as candidate genes in our study. One possible explanation for that is that we do not have power to detect those loci since we performed a very conservative search (i.e., the locus must be an outlier in at least 3 populations). However, it is possible that some of these loci are divergent in individual populations as show in the $F_{ST}$ summary table (i.e. m689 in αA-crystallin for O2O3, m685 in Oca2 for O2O3, Table S3.2, Supplementary material).

*Relationships among population genetic parameters: what outlier loci reveal*

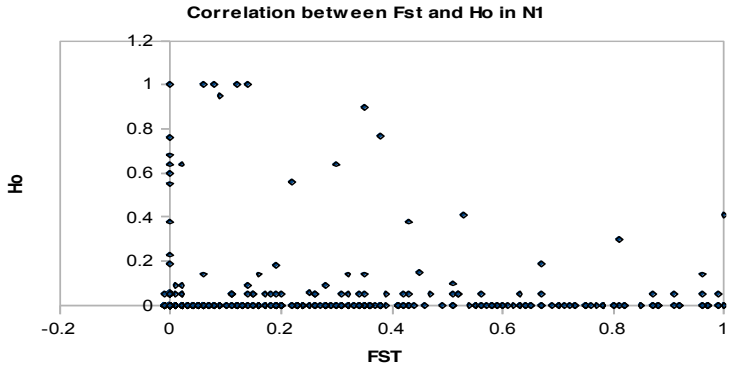
In order to have a better understanding of the highly differentiated loci (outliers) we employed relationship between population genetic parameters per single locus that could give us better insights into outlier behavior. Decreased polymorphism ($H_o$) was observed in a majority of the cave populations, excluding those admixed with the surface (N3* and O8*). Correlation between observed heterozygosities and estimated $F_{ST}$ in each population showed that a majority of the diversified loci (high $F_{ST}$) in old populations are clustered around very low heterozygosities. Also, few loci showed high heterozygosity levels in both new (Figure 3.8. A and B) and old cave populations (Figure 3.8. D, E and F). Compared to isolated cave populations, admixed cave populations (Figure 3.8 C and G) showed shifted distribution of $F_{ST}$ – $H_o$ correlations towards higher $H_o$ and lower $F_{ST}$ values. We did not observe any significant correlations between these two measures in any cave populations; this suggests that the identified $F_{ST}$ outlier loci are not artifacts of overall reduced heterozygosities across the genome.

Finally, to draw some conclusions about the causes of the outlier behavior other than natural selection we performed an HWE test to explore the possibilities of disassortive mating or Wahluand effect. The highest number of the SNPs out of HWE was observed in three populations (N3*, O8* and O2O3 ~ 8%), followed by one surface population (SN1, 7%) while few SNPs were out of the HWE in the other populations (Figure 3.10).

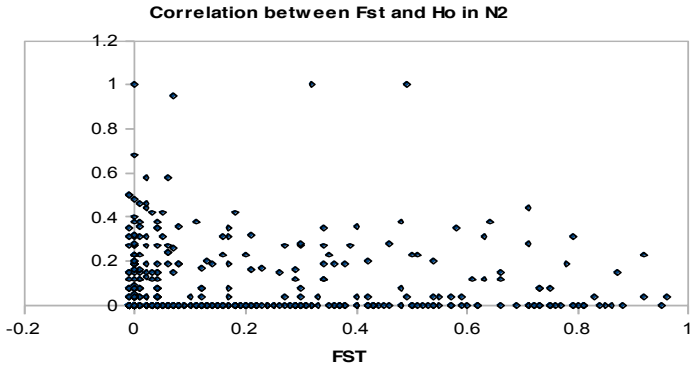
In order to understand relationship of “out of HWE” loci to $F_{ST}$ across loci, we plot the correlation of $F_{IS}$ and $F_{ST}$ for each population at each polymorphic locus. In the cave populations where $F_{IS}$ could be calculated (only for polymorphic loci) we have observed rather extreme values, ether $F_{IS} = 1$ or $F_{IS} = 0$. The intermediate observations were mostly present in admixed populations (N3*, O8*) and in the old population (O4O6). Again, we did not see significant correlations between $F_{ST}$ and $F_{IS}$. However, the $F_{IS}$ in admixed

populations were elevated suggesting that these might be due to the Walhund effect.

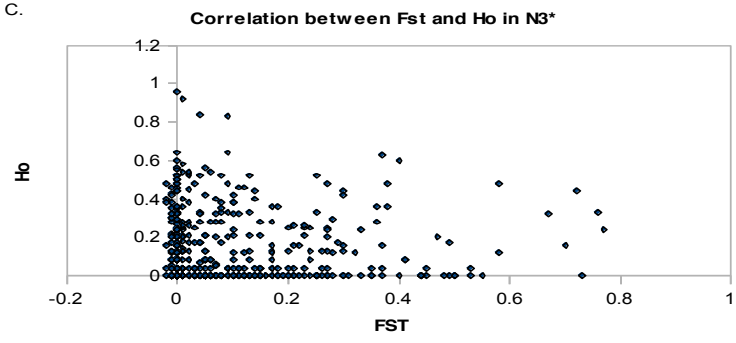
A.


Correlation between Fst and Ho in N1
Ho
FST
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.2
0.4
0.6
0.8
1

B.


Correlation between Fst and Ho in N2
Ho
FST
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.2
0.4
0.6
0.8
1

C.


Correlation between Fst and Ho in N3*
Ho
FST
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.2
0.4
0.6
0.8
1

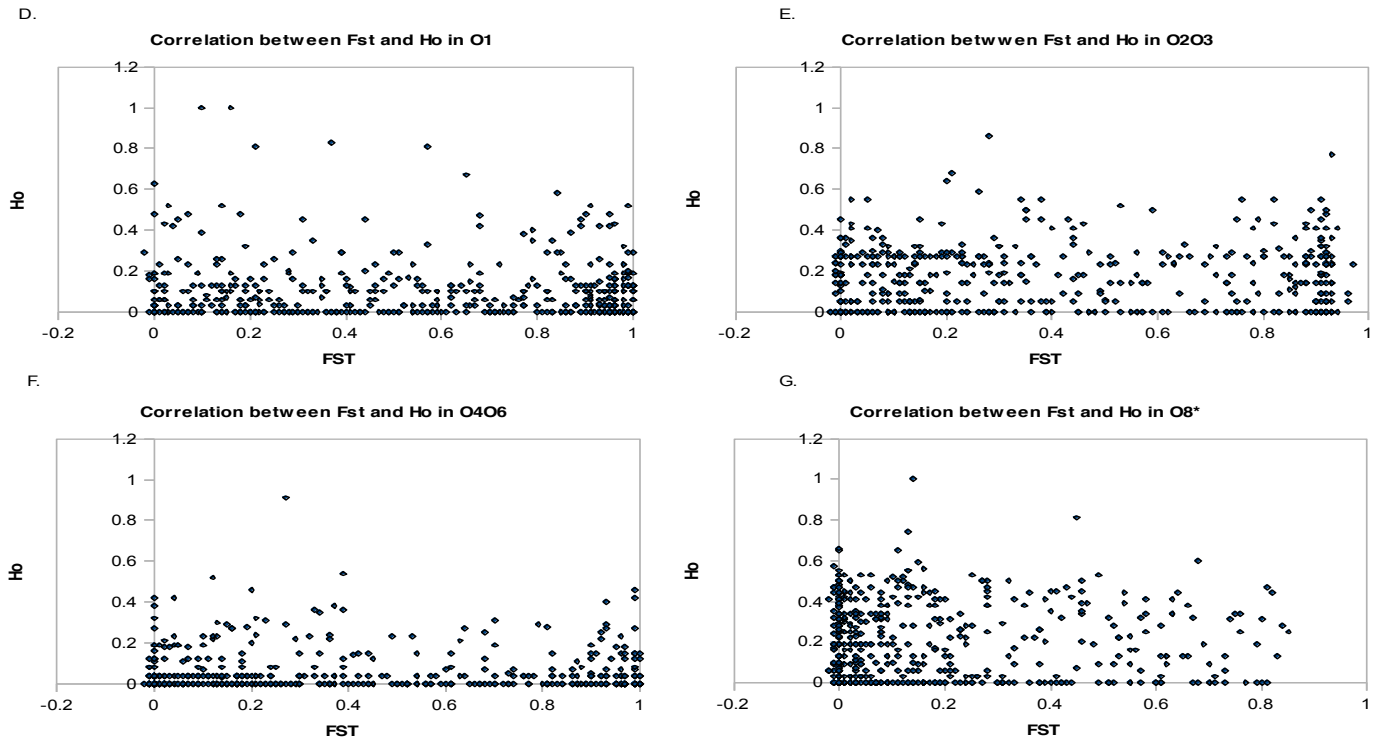
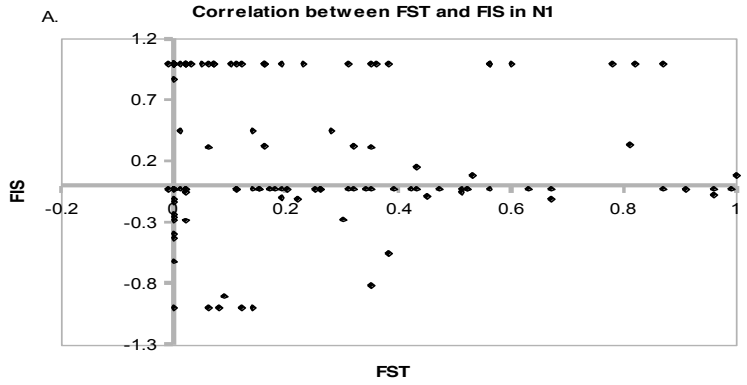
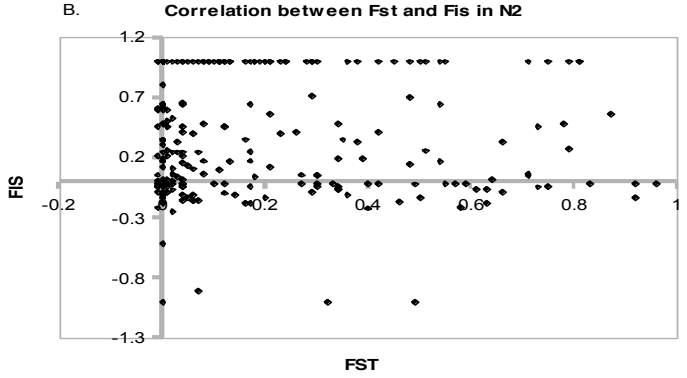
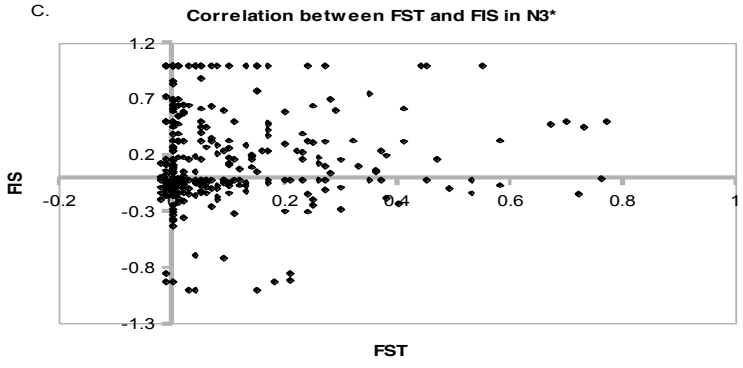



Figure 3.8. Correlation between single locus $F_{ST}$ estimates from outlier loci test *vs.* observed heterozygosities ($H_o$) in multiple cave-surface comparisons. $H_o$ is represented as a function of $F_{ST}$ for all the population pairs and represents estimated $F_{ST}$ from outlier and $H_o$ from single cave population. A. N1 *vs.* surface (SN1, SN2), B. N2 *vs.* surface (SN1, SN2). C. N3* *vs.* surface (SN1, SN2), D. O1 *vs.* surface (SN1, SN2), E. O2O3 *vs.* surface (SN1, SN2), F. O4O6 *vs.* surface (SN1, SN2), G. O8* *vs.* surface (SN1, SN2).


A.
Correlation between FST and FIS in N1
FIS
FST
1.2
0.7
0.2
-0.3
-0.8
-1.3
-0.2
0
0.2
0.4
0.6
0.8
1


B.
Correlation between Fst and Fis in N2
FIS
FST
1.2
0.7
0.2
-0.3
-0.8
-1.3
-0.2
0.2
0.4
0.6
0.8
1


C.
Correlation between FST and FIS in N3*
FIS
FST
1.2
0.7
0.2
-0.3
-0.8
-1.3
-0.2
0.2
0.4
0.6
0.8
1

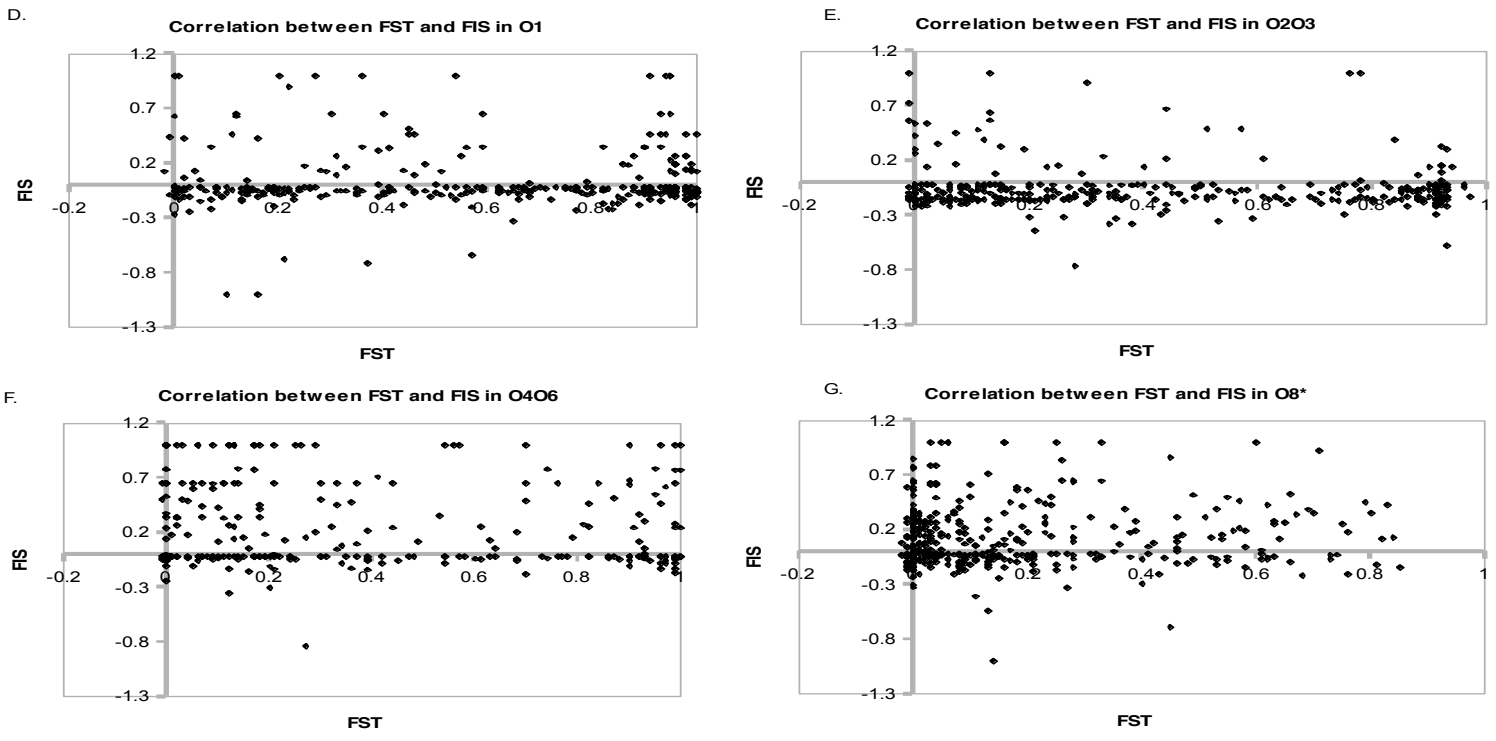



Figure 3.9. Correlation between single locus $F_{ST}$ estimates from outlier loci test *vs.*$F_{IS}$ (inbreeding coefficient) in multiple cave-surface comparisons. $F_{IS}$ is represented as a function of $F_{ST}$ for all the population pairs and represents estimated $F_{ST}$ from outlier and FIS from single cave population. A. N1 *vs.* surface (SN1, SN2), B. N2 *vs.* surface (SN1, SN2). C. N3* *vs.* surface (SN1, SN2), D. O1 *vs.* surface (SN1, SN2), E. O2O3 *vs.* surface (SN1, SN2), F. O4O6 *vs.* surface (SN1, SN2), G. O8* *vs.* surface (SN1, SN2).

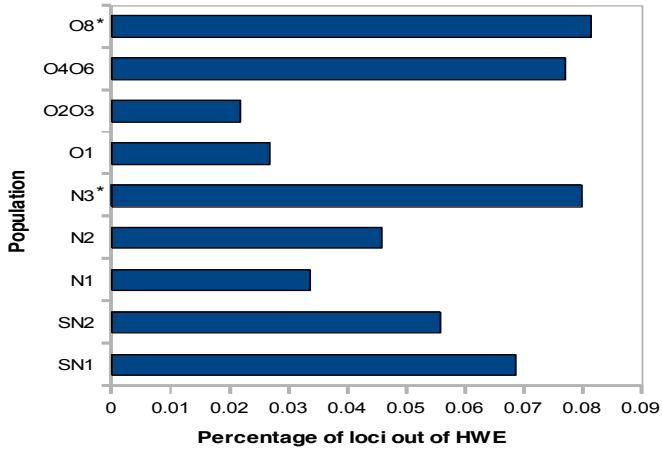



Figure 3.10. Proportion of the loci out of HW equilibrium. Deviations from Hardy-Weinberg Equilibrium (HWE) were estimated per each marker in each population using a Fisher exact test and Bonferroni correction was applied for multiple testing corrections. X-axis is representing the ratio of the HW out of equilibrium and the total number of loci on which the test was performed for each single population. (Note: many loci are monomorphic in certain cave populations (i.e. N1), thus the proportion of the loci out of HW equilibrium is scaled by the number of polymorphic loci). Asterisk represents admixed populations.

### Linkage disequilibrium analysis

In order to understand how evolution in the cave environment structures the population at multiple loci we explored linkage disequilibrium decay (LD) across multiple natural populations. We estimated LD between the markers in each linkage group. Due to the absence of a genome sequence for the *Astyanax* genome; positions of SNPs were based on genetic distances derived from the linkage map. We had sufficient marker densities to perform a LD analysis in only some of the linkage groups (LG1, LG1`, LG7, LG12, LG25) or within BACs. However, with the exclusion of markers with fixed alleles and markers with low minor allele frequencies (MAF < 5%), the numbers of informative markers with known position in the linkage map were extremely low in some populations (N1, N2, O4O6, Figure 3.11). In the other populations, with the higher number of markers left after this quality control we observed only a few, very scattered, markers in each linkage groups. Thus, we are

showing only the overall LD patterns. LD is shown as an $r^2$ estimate within each linkage group as a function of genetic distance in one old cave population (O1), one new cave (N2) population and a surface population (SN1). We were able to observe the general pattern of LD in surface and cave populations (Figure 3.12). Typically, the fraction of marker pairs with high LD does not seem to decreases very fast with distance in cave populations, while surface populations show very low LD overall.

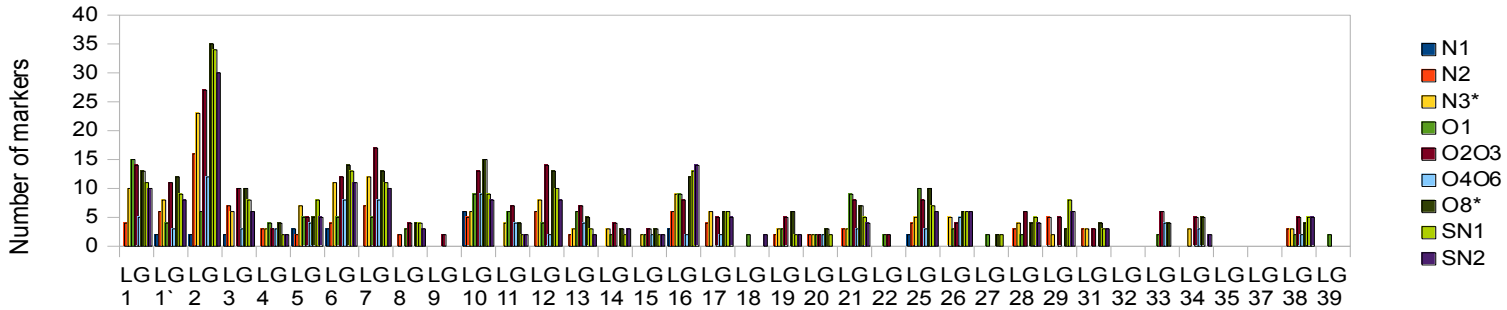



Figure 3.11. Observed marker numbers of markers per each population and linkage group. X-axis is divided in the LG classes and number of markers is shown in each class for each populations. Populations are represented as color-coded bars. Y-axis represents number of markers with MAF > 5% as calculated per each population.

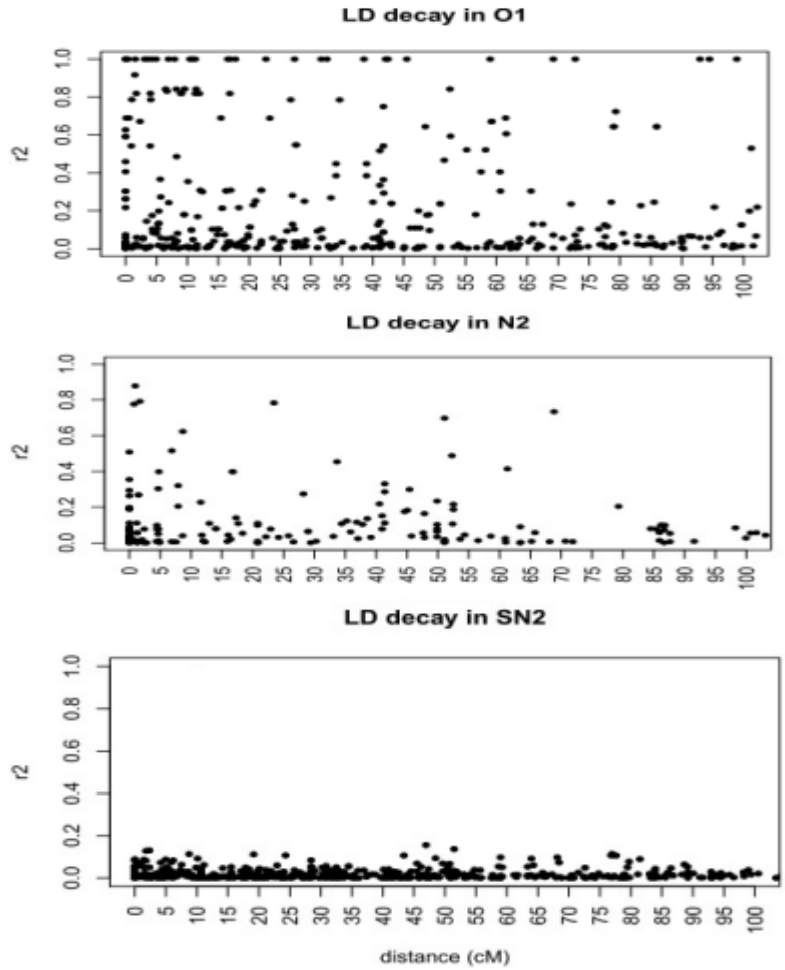



Figure 3.12. LD versus physical distance between SNPs for three population panels: Surface (SN1), old population (O1) and new population (N2). For all three panels, each dot represents pairwise $r^2$ between markers within each linkage group. Values for each linkage group are plotted separately. All the markers that are considered are present in MAF > 5% in the individual populations.

## HAPLOTYPE DIVERSITY

*Outliers and QTL regions*

In order to understand the potential biological significance of the detected outlier loci under the hierarchical model, we further examined differentiation of those loci across different populations, linkage groups and QTL regions. For

that purpose we used the information from our integrated linkage map. We removed microsatellite loci from the map and all the SNP loci were sorted by their linkage groups and positions. In addition, markers that belong to the same BAC clone or the candidate genes that were not mappable (because meiosis for only one of the many markers from that region was informative) were also assigned to the same position at the map as those already placed on the map. In order to test the hypothesis that more outlier markers would be present within QTL region markers were divided in two groups (inside *vs*. outside the QTL region). We saw no significant differences in the number of markers that were inside the QTL region *vs.* number of markers outside the QTL region in a majority of the populations. Significant differences were observed in only two populations that had higher count of markers outside the QTL region ($p<0.01$, O1 and O4O6) (Figure 3.13). We summarize only those loci that are present inside the QTL region and exhibit significance in three or more populations, either within group of origin (new/old) or across the groups (Table 3.4). This conservative approach was taken because some of the loci that have an outlier's behavior might be unreliable due to possible errors in the outlier detection methods. For example, there could be discrepancies when the number of immigrants is unequal between the populations, which are the case with some of our populations (i.e. migration between surface population and O8* cave population as shown in Chapter 2). We identified a total of 80 loci that matched the above-mentioned criteria as summarized in Table 3.4 and those markers were explored in greater detail. Forty-four loci were significant outliers in three or more populations across the comparison between new and old populations. In addition to that we identified 33 loci that were present in at least three populations in only one of the colonization event (Table 3.4.)

| QTL presence | POSITION | LINKAGE GROUP | MARKER | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | MAF>0.05% IN SURFACE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RelEye | 80.74 | LG1 | m164 | 0.08 | **0.04** | -1.00 | 0.05 | **0.02** | **0.02** | **0.04** | - |
| | 91.16 | LG1 | m165 | -1.00 | -1.00 | -1.00 | 0.06 | **0.03** | -1.00 | -1.00 | - |
| | 92.37 | LG1 | m518 | -1.00 | -1.00 | **0.04** | 0.04 | **0.02** | 0.06 | 0.02 | - |
| RelEye | 129.50 | BAC24_LG1' | m669 | **0.02** | **0.01** | 0.08 | **0.00** | **0.03** | **0.01** | **0.00** | - |
| | 129.50 | LG1' | m670 | 0.21 | 0.13 | 0.27 | 0.40 | 0.45 | 0.36 | 0.14 | + |
| | 129.50 | LG1' | m663 | 0.47 | 0.40 | **0.00** | 0.45 | 0.47 | 0.45 | 0.29 | + |
| | 129.50 | LG1' | m660 | 0.41 | 0.18 | 0.38 | 0.34 | 0.38 | 0.34 | **0.02** | + |
| | 129.50 | LG1' | m661 | 0.19 | 0.11 | 0.26 | **0.04** | **0.02** | 0.11 | **0.01** | - |
| | 129.50 | LG1' | m662 | 0.21 | 0.39 | 0.35 | **0.01** | 0.30 | 0.43 | 0.12 | + |
| | 129.50 | LG1' | m665 | 0.23 | 0.26 | 0.26 | 0.44 | 0.41 | 0.37 | 0.19 | + |
| | 130.50 | LG1' | m666 | 0.22 | 0.27 | 0.25 | 0.09 | **0.04** | 0.06 | **0.01** | - |
| | 130.60 | LG1' | m667 | 0.44 | 0.04 | 0.10 | 0.45 | 0.24 | 0.36 | 0.28 | + |
| MelA | 53.28 | BAC1_LG2 | m565 | **0.01** | 0.32 | 0.25 | **0.03** | **0.04** | 0.06 | **0.03** | - |
| | 53.28 | LG2 | m566 | 0.26 | 0.19 | 0.15 | 0.42 | 0.47 | 0.35 | 0.28 | + |
| | 53.29 | LG2 | m564 | -1.00 | 0.05 | -1.00 | -1.00 | -1.00 | **0.04** | **0.02** | - |
| | 53.30 | LG2 | m567 | 0.44 | **0.03** | 0.19 | 0.42 | 0.28 | 0.36 | **0.02** | + |
| | 53.31 | LG2 | m568 | 0.42 | **0.02** | 0.19 | 0.41 | 0.15 | 0.34 | **0.01** | + |
| | 53.38 | LG2 | m569 | 0.22 | 0.27 | 0.25 | 0.33 | 0.22 | 0.26 | 0.35 | - |
| | 53.39 | LG2 | m587 | 0.22 | 0.26 | 0.25 | 0.33 | 0.22 | 0.26 | 0.34 | - |
| | 53.40 | LG2 | m588 | 0.46 | 0.37 | 0.30 | 0.21 | 0.30 | 0.34 | 0.23 | + |
| | 53.41 | LG2 | m589 | 0.42 | 0.36 | 0.23 | 0.47 | 0.26 | 0.47 | 0.41 | + |
| | 53.42 | LG2 | m570 | 0.45 | **0.03** | 0.22 | 0.40 | 0.29 | 0.06 | 0.28 | + |
| | 53.43 | LG2 | m571 | 0.42 | 0.36 | 0.23 | 0.47 | 0.26 | 0.47 | 0.41 | + |
| | 53.44 | LG2 | m572 | 0.39 | 0.33 | 0.20 | 0.40 | 0.33 | 0.41 | 0.39 | + |
| | 53.45 | LG2 | m573 | 0.31 | 0.18 | **0.02** | 0.41 | 0.30 | 0.33 | 0.33 | + |
| | 53.46 | LG2 | m574 | 0.30 | 0.36 | 0.45 | 0.28 | 0.29 | 0.27 | **0.02** | + |
| | 53.48 | LG2 | m577 | 0.35 | 0.27 | 0.35 | 0.47 | 0.26 | **0.01** | 0.28 | + |
| | 53.49 | LG2 | m579 | 0.32 | 0.37 | 0.38 | 0.45 | 0.32 | 0.07 | **0.00** | + |
| | 53.53 | LG2 | m584 | 0.18 | 0.24 | 0.40 | 0.48 | 0.46 | 0.22 | 0.12 | + |
| | 53.54 | LG2 | m585 | 0.18 | 0.24 | 0.46 | 0.48 | 0.46 | 0.22 | 0.12 | + |
| | 53.55 | LG2 | m591 | **0.03** | 0.17 | 0.07 | 0.17 | 0.17 | 0.14 | **0.02** | + |
| | 53.56 | LG2 | m592 | 0.38 | 0.16 | 0.36 | 0.43 | 0.09 | 0.33 | 0.10 | + |
| BAC10 | 100.06 | LG2 | m545 | 0.42 | 0.37 | 0.25 | 0.44 | 0.43 | 0.45 | 0.18 | + |
| | 100.07 | LG2 | m543 | 0.47 | 0.40 | 0.30 | 0.45 | 0.47 | 0.47 | 0.16 | + |
| | 100.08 | LG2 | m544 | 0.29 | 0.16 | 0.16 | 0.41 | 0.32 | 0.38 | 0.29 | + |
| | 100.09 | LG2 | m546 | 0.40 | 0.12 | 0.15 | 0.25 | 0.20 | 0.18 | 0.19 | + |
| | 100.10 | LG2 | m547 | 0.46 | 0.43 | 0.34 | 0.13 | 0.13 | 0.09 | 0.08 | + |
| | 100.12 | LG2 | m549 | 0.37 | 0.18 | 0.22 | 0.36 | 0.26 | 0.28 | 0.27 | + |
| | 100.13 | LG2 | m553 | 0.21 | 0.09 | **0.02** | 0.11 | 0.13 | 0.25 | 0.08 | + |
| | 100.14 | LG2 | m550 | 0.22 | 0.28 | 0.42 | 0.41 | 0.33 | 0.29 | 0.43 | + |
| | 100.16 | LG2 | m552 | 0.42 | 0.35 | 0.37 | 0.48 | 0.37 | 0.35 | 0.27 | + |
| | 100.17 | LG2 | m557 | 0.48 | 0.06 | 0.37 | 0.14 | 0.31 | 0.43 | **0.01** | + |
| | 100.23 | LG2 | m563 | 0.22 | 0.26 | 0.25 | 0.33 | 0.22 | 0.26 | 0.34 | - |
| | 100.24 | LG2 | m554 | 0.46 | 0.36 | 0.34 | 0.43 | 0.22 | 0.34 | 0.11 | + |
| | 100.26 | LG2 | m556 | 0.25 | 0.29 | **0.02** | 0.21 | 0.17 | 0.23 | 0.11 | + |
| LensL | 113.18 | LG3 | m193 | 0.22 | 0.27 | 0.25 | **0.04** | 0.05 | 0.23 | 0.35 | - |
| | 116.91 | LG3 | m446 | 0.23 | 0.22 | 0.12 | 0.23 | 0.22 | 0.23 | 0.14 | + |
| | 119.16 | LG3 | m15 | -1.00 | -1.00 | -1.00 | **0.00** | 0.08 | 0.11 | 0.12 | - |
| | 119.16 | LG3 | m538 | 0.32 | 0.45 | 0.34 | 0.49 | 0.37 | 0.44 | 0.40 | + |
| | 119.36 | LG3 | m17 | -1.00 | -1.00 | -1.00 | **0.00** | 0.05 | **0.00** | **0.01** | - |
| | 122.24 | LG3 | m703 | **0.01** | **0.01** | **0.00** | 0.09 | **0.04** | **0.02** | **0.01** | - |
| | 122.25 | LG3 | m702 | 0.26 | **0.04** | 0.07 | 0.39 | **0.01** | 0.31 | 0.42 | + |
| ResidLen | 7.99 | LG10 | m30 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | -1.00 | -1.00 | - |
| | 12.38 | LG10 | m284 | 0.09 | 0.05 | -1.00 | **0.02** | **0.02** | 0.06 | **0.02** | - |
| | 14.52 | LG10 | m286 | 0.09 | 0.05 | -1.00 | **0.02** | **0.01** | 0.06 | **0.01** | - |
| | 15.63 | LG10 | m276 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | -1.00 | -1.00 | - |
| | 15.86 | LG10 | m605 | 0.14 | 0.41 | **0.03** | 0.16 | 0.22 | 0.48 | 0.19 | + |
| | 15.87 | LG10 | m606 | 0.33 | 0.41 | 0.07 | 0.30 | 0.32 | 0.30 | 0.41 | + |
| | 15.88 | LG10 | m598 | 0.37 | 0.27 | 0.39 | **0.03** | 0.24 | 0.14 | 0.19 | + |
| | 15.89 | LG10 | m599 | 0.12 | 0.30 | 0.10 | 0.07 | **0.04** | 0.25 | 0.11 | + |
| | 15.91 | LG10 | m601 | 0.26 | **0.03** | **0.04** | 0.34 | 0.27 | 0.24 | 0.45 | + |
| | 15.97 | LG10 | m612 | 0.36 | 0.18 | 0.12 | 0.25 | 0.46 | **0.02** | 0.35 | + |
| | 15.98 | LG10 | m613 | 0.32 | 0.37 | 0.40 | 0.27 | 0.31 | **0.02** | 0.25 | + |
| | 16.86 | LG10 | m514 | 0.30 | 0.41 | 0.33 | **0.01** | **0.03** | 0.16 | **0.03** | + |
| | 16.89 | LG10 | m515 | 0.41 | 0.46 | 0.35 | 0.39 | **0.02** | **0.03** | **0.04** | - |
| | 16.90 | LG10 | m620 | **0.03** | **0.04** | 0.44 | 0.41 | 0.37 | 0.36 | 0.46 | + |
| | 16.96 | LG10 | GHSNP1 | 0.47 | 0.15 | 0.16 | 0.18 | 0.14 | 0.15 | 0.46 | - |
| | 17.78 | LG10 | m18 | 0.18 | 0.23 | 0.14 | 0.30 | 0.19 | 0.23 | 0.09 | + |
| | 18.69 | LG10 | m286 | 0.08 | **0.04** | -1.00 | 0.13 | 0.04 | **0.02** | 0.17 | - |
| | 20.42 | LG10 | m290 | 0.25 | 0.14 | 0.25 | 0.13 | 0.05 | 0.06 | 0.07 | - |
| | 22.77 | LG10 | m294 | 0.18 | **0.04** | 0.37 | 0.38 | 0.27 | 0.30 | 0.04 | + |
| AASens | 10.99 | LG11 | m311 | 0.08 | **0.04** | -1.00 | **0.02** | **0.02** | **0.04** | **0.02** | - |
| | 12.30 | LG11 | m308 | 0.05 | 0.09 | 0.09 | 0.05 | 0.06 | 0.09 | 0.08 | - |

| QTL presence | POSITION | LINKAGE GROUP | MARKER | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | MAF>0.05% IN SURFACE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 22.50 | LG12 | m673 | 0.37 | 0.42 | 0.43 | 0.48 | 0.36 | 0.39 | **0.03** | + |
| | 22.50 | LG12 | m677 | 0.30 | **0.02** | 0.43 | 0.30 | 0.30 | 0.29 | 0.44 | + |
| | 22.50 | LG12 | m680 | 0.14 | 0.05 | 0.24 | 0.26 | 0.39 | 0.42 | 0.44 | + |
| RelEye & Mel | 22.50 | LG12 | m682 | -1.00 | -1.00 | -1.00 | **0.02** | **0.02** | **0.04** | **0.02** | - |
| | 22.50 | LG12 | m685 | -1.00 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | -1.00 | - |
| | 32.27 | LG12 | m312 | 0.12 | 0.20 | **0.04** | 0.22 | 0.22 | 0.20 | 0.15 | + |
| | 34.00 | LG12 | m296 | 0.30 | 0.45 | 0.19 | 0.42 | 0.48 | 0.38 | 0.25 | + |
| | 38.82 | LG12 | m497 | 0.15 | 0.18 | **0.03** | 0.06 | 0.15 | 0.17 | 0.26 | + |
| | 44.75 | LG12 | m466 | 0.28 | **0.02** | 0.44 | 0.21 | 0.18 | 0.21 | 0.40 | + |
| | 29.24 | LG14 | m309 | -1.00 | -1.00 | -1.00 | **0.02** | 0.09 | -1.00 | -1.00 | - |
| RelEye | 35.51 | LG14 | m315 | **0.03** | **0.04** | **0.04** | 0.12 | 0.12 | 0.07 | 0.16 | + |
| | 38.50 | LG14 | m477 | 0.27 | 0.27 | 0.48 | **0.02** | 0.29 | 0.37 | 0.29 | + |
| | 1.62 | LG16 | m155 | **0.01** | 0.11 | 0.24 | 0.22 | 0.23 | 0.24 | 0.36 | - |
| | 1.62 | LG16 | m159 | 0.22 | 0.14 | 0.25 | 0.10 | 0.06 | **0.02** | 0.21 | - |
| | 5.90 | LG16 | m454 | -1.00 | -1.00 | **0.04** | **0.04** | 0.01 | 0.20 | -1.00 | - |
| | 10.39 | LG16 | m489 | 0.37 | 0.32 | 0.19 | 0.08 | 0.39 | 0.42 | 0.26 | + |
| | 16.43 | LG16 | m170 | 0.46 | **0.03** | 0.18 | 0.05 | 0.17 | 0.30 | 0.44 | + |
| | 20.93 | LG16 | m626 | 0.42 | 0.32 | 0.26 | 0.48 | 0.35 | 0.35 | 0.40 | - |
| | 20.94 | LG16 | m627 | 0.32 | 0.37 | 0.37 | 0.46 | 0.32 | 0.35 | 0.15 | + |
| | 20.95 | LG16 | m628 | 0.34 | 0.05 | 0.24 | 0.40 | 0.35 | 0.49 | 0.24 | + |
| | 20.96 | LG16 | m629 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | - |
| | 20.97 | LG16 | m630 | 0.27 | 0.19 | 0.23 | 0.13 | 0.27 | 0.22 | 0.11 | + |
| AASens (BAC 6) | 21.02 | LG16 | m637 | **0.02** | **0.00** | 0.10 | 0.13 | **0.03** | 0.12 | **0.02** | - |
| | 21.03 | LG16 | m639 | 0.36 | 0.16 | 0.24 | 0.34 | 0.41 | 0.35 | 0.22 | + |
| | 21.05 | LG16 | m642 | 0.46 | 0.11 | 0.18 | 0.37 | 0.33 | 0.32 | 0.15 | + |
| | 21.09 | LG16 | m632 | 0.42 | 0.18 | **0.00** | **0.04** | **0.03** | 0.33 | 0.07 | - |
| | 21.11 | LG16 | m634 | 0.46 | 0.40 | **0.03** | 0.38 | 0.33 | 0.39 | **0.01** | + |
| | 21.12 | LG16 | m638 | 0.39 | 0.45 | 0.18 | 0.34 | 0.36 | 0.32 | 0.31 | + |
| | 21.13 | LG16 | m648 | 0.22 | 0.30 | 0.30 | 0.21 | 0.24 | 0.31 | 0.18 | - |
| | 21.17 | LG16 | m652 | 0.32 | 0.25 | 0.14 | 0.45 | 0.38 | 0.41 | 0.42 | + |
| | 21.20 | LG16 | m655 | 0.33 | 0.24 | 0.30 | 0.44 | 0.35 | 0.41 | 0.37 | + |
| | 21.22 | LG16 | m657 | 0.09 | **0.04** | **0.04** | **0.01** | 0.23 | 0.23 | **0.03** | + |
| | 21.23 | LG16 | m658 | 0.46 | 0.05 | 0.46 | 0.45 | 0.43 | 0.22 | 0.05 | + |
| | 21.24 | LG16 | m659 | 0.40 | 0.46 | 0.46 | 0.32 | 0.21 | 0.33 | 0.37 | + |
| | 20.08 | LG25 | m252 | **0.00** | 0.09 | 0.11 | **0.04** | 0.34 | **0.04** | **0.02** | - |
| Reldent | 20.99 | LG25 | m257 | -1.00 | -1.00 | -1.00 | **0.02** | 0.10 | -1.00 | -1.00 | - |
| | 28.63 | LG25 | m476 | 0.29 | 0.19 | 0.12 | 0.39 | 0.41 | 0.39 | 0.26 | + |
| | 37.80 | LG25 | m10 | -1.00 | -1.00 | -1.00 | **0.03** | **0.04** | **0.03** | 0.12 | - |
| | 0.81 | LG28 | m224 | 0.05 | 0.09 | 0.09 | 0.07 | 0.05 | 0.09 | 0.21 | - |
| COND | 11.94 | LG28 | m219 | 0.17 | **0.01** | 0.25 | 0.13 | **0.04** | 0.06 | **0.04** | - |
| | 16.22 | LG28 | m221 | 0.16 | **0.02** | 0.07 | 0.06 | 0.09 | 0.06 | 0.30 | + |
| | 26.53 | LG28 | m222 | -1.00 | -1.00 | -1.00 | **0.00** | **0.03** | -1.00 | -1.00 | - |
| Wtloss | 29.01 | LG28 | m708 | 0.27 | 0.13 | 0.23 | 0.39 | 0.47 | 0.35 | 0.22 | + |
| | 29.02 | LG28 | m709 | 0.27 | 0.17 | 0.42 | 0.39 | 0.15 | 0.35 | 0.23 | + |
| | ^ | LG28 | m457 | **0.04** | 0.12 | **0.02** | 0.15 | 0.30 | **0.02** | 0.08 | - |
| | 14.50 | LG29 | m688 | 0.26 | 0.22 | 0.30 | 0.40 | 0.09 | 0.31 | 0.14 | + |
| | 14.54 | LG29 | m696 | 0.46 | 0.43 | 0.20 | 0.47 | 0.47 | 0.47 | 0.43 | + |
| | 14.55 | LG29 | m686 | 0.45 | 0.46 | 0.25 | 0.44 | 0.45 | 0.43 | 0.35 | + |
| RelEye | 14.56 | LG29 | m690 | 0.45 | 0.38 | 0.25 | 0.15 | 0.45 | 0.49 | 0.37 | + |
| | 14.51 | LG29 | m689 | 0.44 | 0.34 | 0.45 | 0.40 | **0.02** | 0.28 | 0.06 | + |
| | 14.52 | LG29 | m695 | 0.18 | 0.05 | 0.36 | 0.37 | 0.47 | 0.44 | 0.35 | + |
| | 14.51 | LG29 | m693 | 0.25 | 0.39 | 0.12 | 0.47 | 0.37 | 0.38 | 0.21 | + |
| | 14.53 | LG29 | m694 | 0.17 | 0.08 | 0.15 | 0.35 | 0.49 | 0.43 | 0.34 | + |
| | 2.77 | LG33 | m443 | -1.00 | -1.00 | -1.00 | **0.01** | **0.02** | **0.00** | **0.02** | - |
| | 3.99 | LG33 | m452 | -1.00 | -1.00 | -1.00 | **0.02** | **0.04** | **0.02** | **0.01** | - |
| WtLoss & COND | 5.82 | LG33 | m230 | -1.00 | -1.00 | -1.00 | 0.09 | -1.00 | -1.00 | -1.00 | - |
| | 5.82 | LG33 | m232 | 0.19 | 0.17 | **0.03** | 0.32 | 0.33 | 0.28 | 0.32 | + |
| | 5.82 | LG33 | m235 | -1.00 | **0.04** | -1.00 | **0.03** | **0.02** | **0.01** | 0.06 | - |
| | 7.24 | LG33 | m213 | **0.03** | 0.27 | 0.23 | 0.12 | 0.07 | **0.04** | **0.00** | - |
| | 16.89 | LG33 | m231 | 0.05 | 0.09 | 0.09 | 0.05 | 0.06 | 0.06 | 0.22 | - |
| | 6.84 | LG34 | m479 | -1.00 | -1.00 | **0.04** | **0.00** | **0.02** | **0.00** | **0.00** | - |
| | 12.99 | LG34 | m237 | 0.06 | 0.08 | 0.11 | 0.05 | **0.03** | 0.05 | 0.25 | - |
| LEN MEL_E MEL A | 14.72 | LG34 | m239 | -1.00 | **0.04** | 0.15 | 0.13 | **0.03** | 0.11 | 0.13 | - |
| | 15.33 | LG34 | m500 | 0.09 | 0.06 | 0.13 | **0.01** | 0.07 | 0.06 | 0.06 | - |
| | 20.37 | LG34 | m238 | 0.40 | 0.37 | 0.24 | 0.39 | 0.32 | 0.39 | 0.23 | + |
| | 25.08 | LG34 | m503 | 0.28 | 0.31 | 0.06 | 0.13 | 0.04 | 0.28 | 0.09 | + |
| RelEye | 0.00 | LG37 | m161 | -1.00 | -1.00 | -1.00 | 2.00 | -1.00 | -1.00 | -1.00 | - |
| | 8.58 | LG37 | m410 | 0.28 | **0.02** | 0.39 | 0.25 | 0.20 | **0.01** | **0.02** | + |
| | 17.14 | LG38 | m210 | 0.32 | 0.41 | 0.38 | 0.06 | 0.36 | **0.02** | 0.41 | + |
| Mel_E | 22.51 | LG38 | m204 | **0.00** | **0.01** | -1.00 | **0.01** | 0.15 | -1.00 | 0.06 | - |
| | 22.51 | LG38 | m209 | 0.10 | 0.05 | 0.07 | 0.32 | 0.09 | 0.21 | 0.07 | + |
| | 24.66 | LG38 | m207 | 0.12 | 0.38 | 0.05 | 0.28 | 0.31 | 0.29 | **0.04** | + |
| | 27.01 | LG38 | m494 | -1.00 | -1.00 | 0.18 | **0.00** | 0.19 | -1.00 | -1.00 | - |

Table 3.4. Summary of significant $F_{ST}$ values assigned to the QTL locus. Only QTL regions and markers that are significant and fall into the QTL region with their surrounding markers are shown. Each marker is labeled as + or – based on its presence in "surface SNPs" (MAF > 5%) or "cave SNPs" (MAF < 5%). -1 value determines that the variance was not possible to calculate since there was no variation between particular population and the surface population. Values in the table represent P-values < 0.05 as identified by coalescent simulations in ARELQUIN 3.5. Red labels represents loci that are detected as significant outliers across old and new caves in at least three populations, while blue labels represent markers that are significant in at least three populations in only new or old group. Light gray blocks represent the block of QTL locus and dark gray-labeled markers represent the markers that were considered in the phasing process.

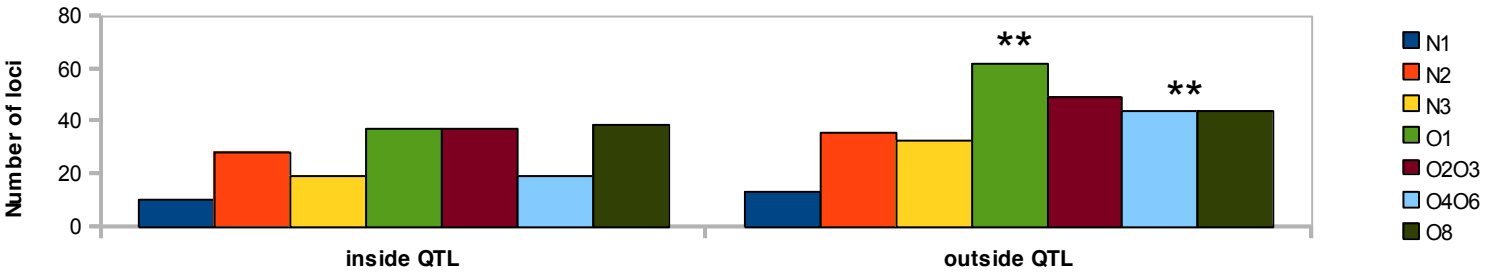



Figure 3.13. Summary of the outliers found inside or outside the population per each population. X-axis divides marker in two groups (out of QTL or inside the QTL region). Y-axis represents the outlier number. Colors are attributed to the different populations. Asterisk represent significant difference ($p < 0.05$) as identified by $\chi^2$ test on the outlier number when comparing outlier counts based on two criteria. ($\chi^2$ = 9.9206, DF = 1, p-value = 0.001634 for O2O3 and $\chi^2$ = 6.3131, DF = 1, p-value = 0.01198 for O1).

*Haplotype phasing and diversity*

We performed analyses of haplotype diversity focusing primarily on the linkage groups and QTL regions where outliers SNPs were detected. We are showing here only those linkage groups in which we observed an interesting pattern while contrasting different cave lineages and surface populations. Haplotype diversity was measured across overlapping windows of 2 SNPs in a given QTL region using measure of effective haplotype number (*he*). Effective number of haplotypes was estimated as $he = 1/\Sigma pi^2$, with *pi* the frequency of haplotype *i* for a total number of *h* haplotypes. *he* was generally lower in cave populations and ranged from 1 to 1.5 in isolated cave populations (N1, N2, O1, O2O3, O4O6), while admixed cave populations showed more diversity (N3* and O8*). In contrast *he* in surface populations ranged from 1.5 to 2 (Figure 3.14). *he* is directly related to haplotype heterozygosity (the probability of two different haplotypes in one individual; *see the Methods section*). Thus, the lower effective number of haplotypes is suggestive of lower haplotype heterozygosities in cave populations in comparison to surface populations. We also estimated percentages of common haplotypes per QTL region in the combined sample and made the following comparisons: new caves *vs.* surface, old caves *vs.* surface and new caves *vs.* old caves. Proportions of common haplotypes were the highest across new caves *vs.* surface comparisons, in all the regions and ranged from 40 to 80%.

We have also identified regions in which the shared haploype was observed among caves from different lineages but not in the surface populations. For example, the first two positions in the overlapping window in LG12 within the QTL for eye and melanin (EyeMel_D) showed common haplotypes between the old and new lineages but there were no haplotypes shared with the surface populations (Figure 3.14.C).

*Haplotype divergence within and between different lineages*

We estimated variance between and within new caves *vs.* surface; old cave *vs.* surface and new caves *vs.* old caves comparisons and observed significant differences across different levels of comparisons (between and within the groups). We explored each QTL region and its associated LOD profile that describes the strength of the marker association with the given trait across the linkage group. All the markers that are present within that LOD score profile were used in phasing individual genotypes into haplotypes and the haplotype divergence across overlapping SNP windows was compared between the population groups.

The melanin QTL (MEL_A) in LG2 showed three outlier markers across new and old populations (Figure 3.14. A. BAC1, outlier markers are designated in red color on the linkage map). We have observed strong divergence of old cave populations from surface populations for five haplotypes in this QTL region (sliding window position 1, 3, 4, 5, 6, Figure 3.14.A.ii. BAC1). Those haplotypes were also shared in a high proportion with the new cave populations (sliding window 1 to 15 and 19 to 22, Figure 3.14.A.i. BAC1) and the significant diversification in old caves *vs.* new caves comparisons was not observed (sliding window position 15 and 19 to 22, Figure 3.14.A.iv. BAC1). However, divergence on the global cave-surface comparison was observed only in the old populations (sliding window positions 1 to 5, Figure 3.14.A.iv. BAC1). Within population variance (Figure 13.4.A.v.) in this QTL showed significant diversification from the surface population that was due to the individual populations within new or old cave groups. This suggests that different caves from the same lineage have different levels of divergence from surface populations (Figure 3.14.A.v, sliding window position 16, 17, 18). To contrast this pattern within the QTL region with the region out of QTL, we also estimated diversification of BAC10 region in LG2. Haplotypes within BAC 10 showed lower level of shared haplotypes when comparing old and new cave

populations and slightly higher haplotype diversities (*he*) than in BAC1 region (Figure.3.14. A.ii., iii. BAC1 *vs.* BAC10 panel). In BAC10 region in LG2 (outside of QTL) diversification of cave *vs.* surface populations was also significant for some haplotypes. Here, we have also observed divergence of both lineages (old and new) from surface population at the first position of the sliding window. However, the overall diversification in BAC10 region is smaller than the one within the QTL region, BAC1 (Figure 3.14. iv., v., BAC1 *vs.* BAC10 panel).

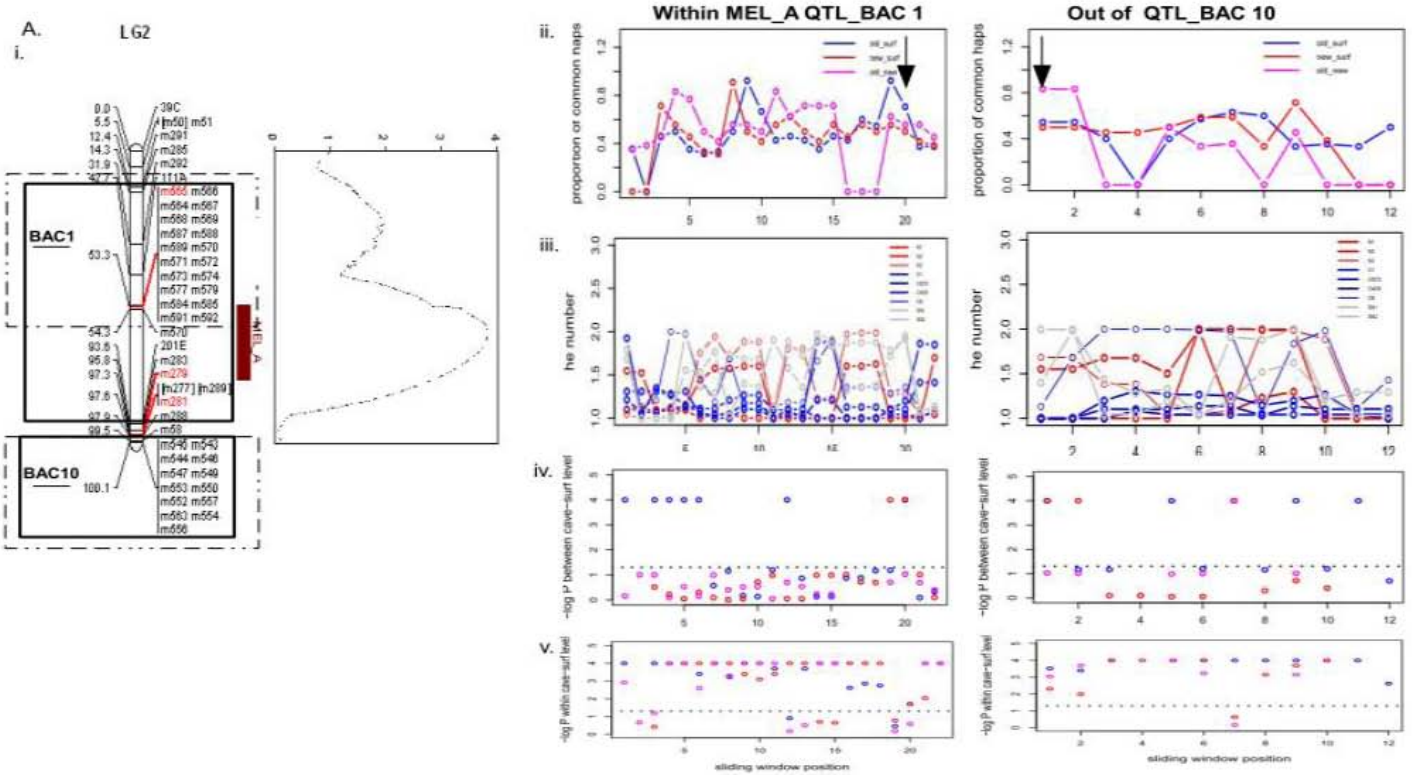
We have detected two significant outlier markers in BACGH region in LG10 (Figure 3.14. B.i.). Also, strong haplotype divergence of the new caves from the surface populations was also detected for the QTL responsible for the length within the BACGH that contains growth hormone gene (Figure 3.14. B.ii.). Our results show significant diversification of new cave populations from the surface populations for multiple haplotypes (Figure 3.14.B.iv., sliding window 14, 15 and 17, 18). Also, majority of the haplotypes in this region were highly differentiated between old caves and new caves (Figure 3.14.B.iv and v).

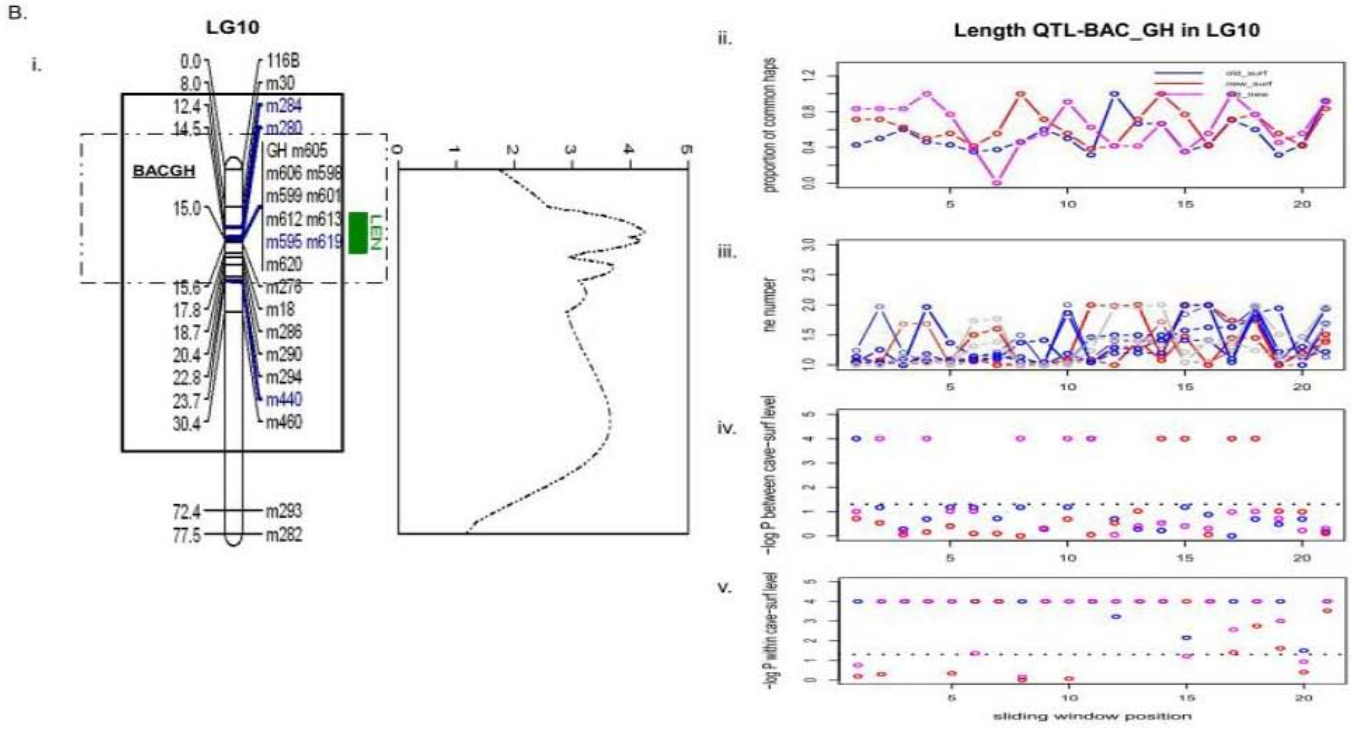
We have further explored diversification in LG12 where the overlapping QTL regions for eye and melanophore numbers were observed. This region contained only one significant outlier (Figure 3.14.C.i.). Here, we detected two haplotypes that were unobserved in surface populations but they were shared in both cave lineages (old and new) (Figure 3.14.C.ii., sliding window positions 1 and 2). Also differentiation between the two groups of cave populations was not significant (Figure 3.14.C.iv., sliding window positions 1 and 2). However, we have observed significant diversification within the new and old lineages suggesting variability in the haplotype frequencies for the individual populations within the lineage (Figure 3.14.C.v., sliding window positions 1 and 2).

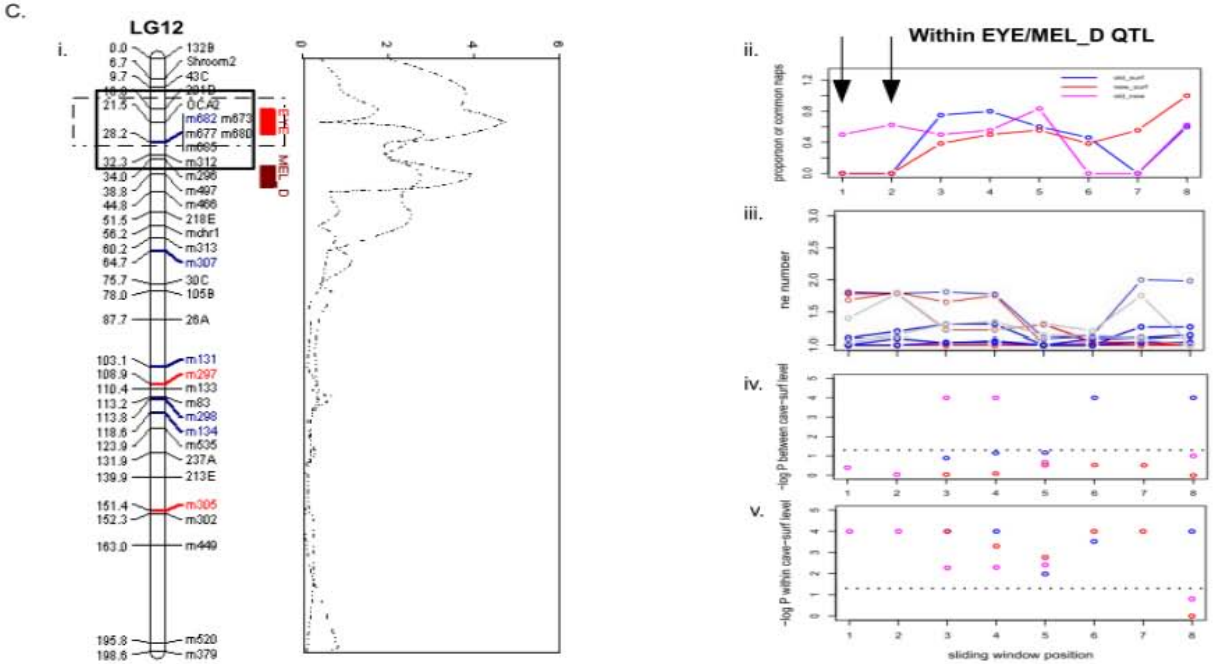
Similarly, both old and new caves shared two haplotypes that differ from the surface population in the QTL for amino-acid sensitivity in LG16 suggesting

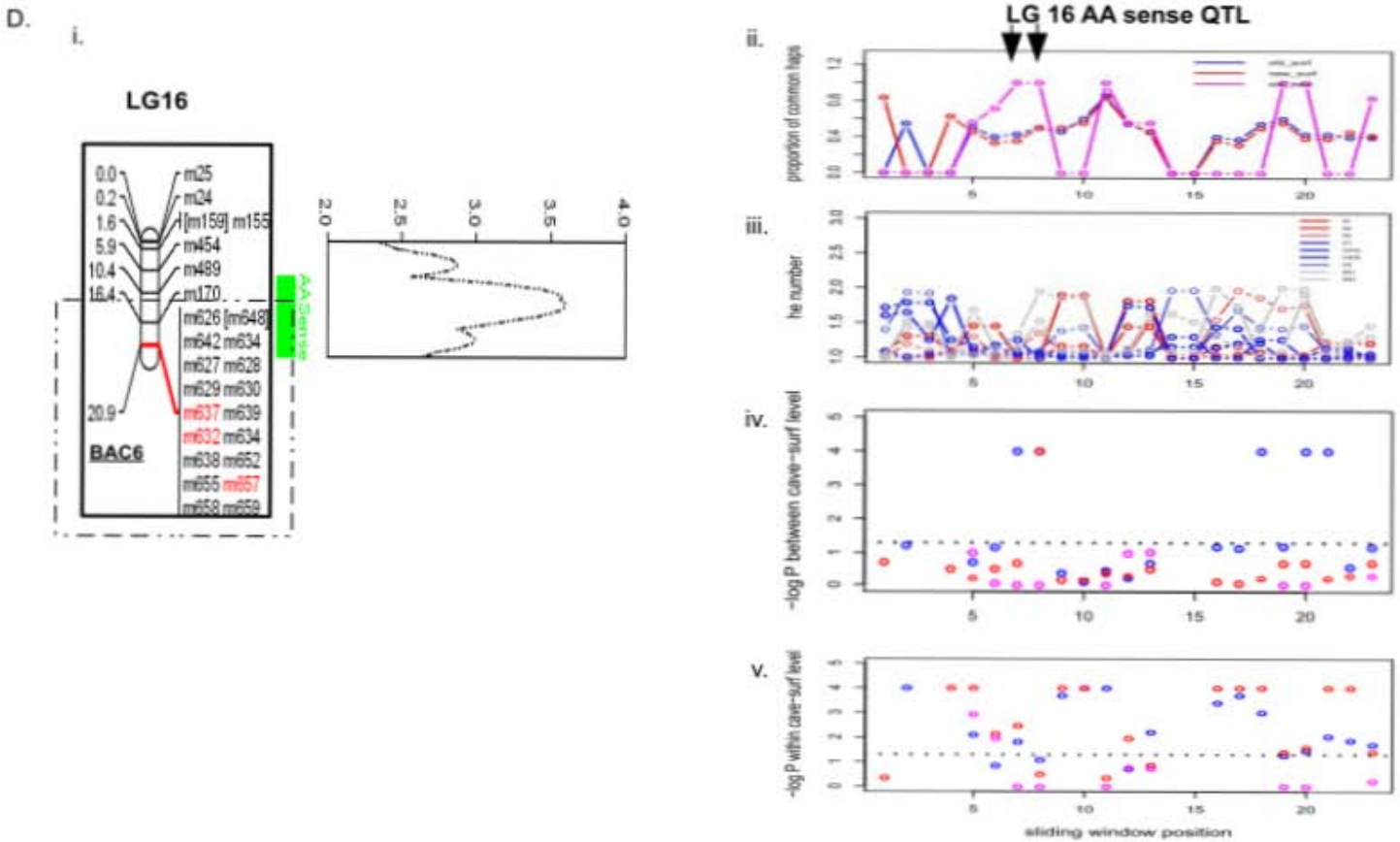
convergent genetic evolution (Fig 3.14.D.ii., iii.,iv., sliding window position 7 and 8). Also, in this QTL we have noticed 3 haplotypes next to each other that were significantly different from the surface populations suggesting parallel diversification in the old lineage (Fig 3.14.D.iv., sliding window position 18, 20 and 21). However, in both of these cases we did not observe significant diversification from all the populations in the individual lineages. That is shown in within the population variance that was significant (Fig 3.14.D.v., sliding window position 7, 8, 18, 20 and 21). These observations in the QTL of LG12 and LG16 are suggestive of possible convergent evolution in those regions, and interestingly divergent haplotypes are not observed in surface population in the region of LG12. These could be because very low frequencies of those haplotypes are present in surface populations, so we cannot observe them. Alternatively, these haplotypes are probably a result of the distant causative marker, which is in high LD with the scanned markers.

In the summary most of the haplotypes are common within each lineage. Surface populations share high level of haplotypes with the new cave populations. Diversification of individual haplotypes from the surface populations sometimes varies within the lineage, which suggests that parallel changes are not always result of diversification.


A.
i.
LG2
BAC1
BAC10
MEL_A
Within MEL_A QTL_BAC 1
Out of QTL_BAC 10
ii.
iii.
iv.
v.
proportion of common haps
he number
-log P between cave−surf level
-log P within cave−surf level
sliding window position


B.
i.
LG10
0.0
8.0
12.4
14.5
15.0
15.6
17.8
18.7
20.4
22.8
23.7
30.4
72.4
77.5
BACGH
116B
m30
m284
m280
GH m605
m606 m598
m599 m601
m612 m613
m595 m619
m620
m276
m18
m286
m290
m294
m440
m460
m293
m282
LEN
0
1
2
3
4
5
ii.
Length QTL-BAC_GH in LG10
proportion of common haps
iii.
ne number
iv.
-log P between cave-surf level
v.
-log P within cave-surf level
sliding window position


C.
i.
LG12
0.0 132B
6.7 Shroom2
9.7 43C
21.5 OCA2
m682 m873
28.2 m677 m880
m685
32.3 m312
34.0 m296
38.8 m497
44.8 m466
51.5 218E
56.2 mchr1
60.2 m313
64.7 m307
75.7 30C
78.0 105B
87.7 26A
103.1 m131
108.9 m297
110.4 m133
113.2 m83
113.8 m298
118.6 m134
123.9 m535
131.9 237A
139.9 213E
151.4 m305
152.3 m302
163.0 m449
195.8 m520
198.6 m379
EYE
MEL_D
0
2
4
6
ii.
Within EYE/MEL_D QTL
proportion of common naps
iii.
ne number
iv.
-log P between cave-surf level
v.
-log P within cave-surf level
sliding window position


D.
i.
LG16
0.0
0.2
1.6
5.9
10.4
16.4
20.9
m25
m24
[m159] m155
m454
m489
m170
m626 [m648]
m642 m634
m627 m628
m629 m630
m637 m639
m632 m634
m638 m652
m655 m657
m658 m659
BAC6
AA sense
2.0
2.5
3.0
3.5
4.0
LG 16 AA sense QTL
ii.
proportion of common haps
iii.
he number
iv.
-log P between cave-surf level
v.
-log P within cave-surf level
5
10
15
20
sliding window position

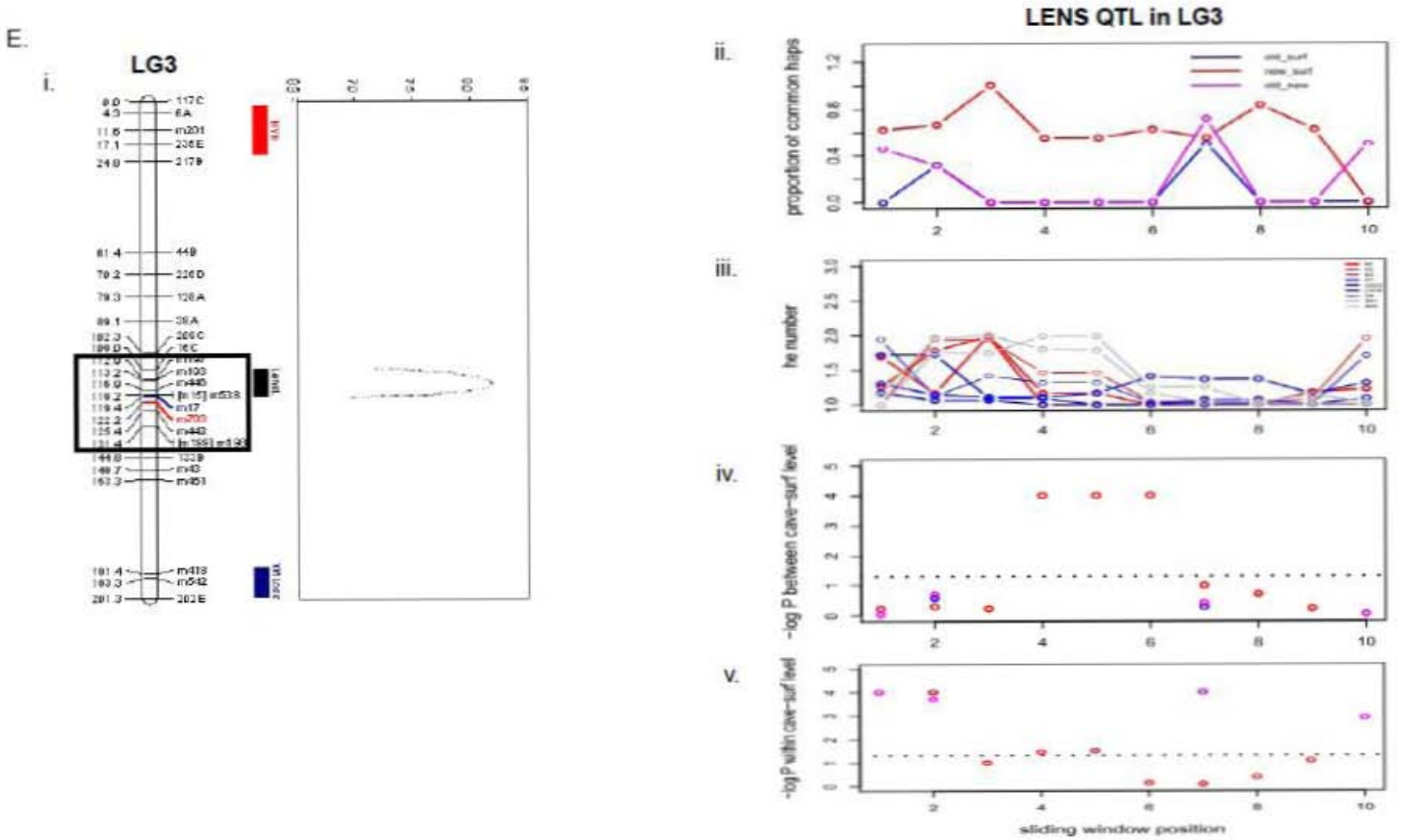

E.
i.
LG3
ii.
LENS QTL in LG3
proportion of common haps
iii.
he number
iv.
-log P between cave-surf level
v.
-log P within cave-surf level
sliding window position

Figure 3.14. Haplotypes and population comparisons of the different QTLs and linkage groups. A. QTL for melanin (MELA) in LG2 and BAC10 region out of QTL B. QTL for length in LG10 and BACGH with the QTL C. QTL for eye and melanophore numbers (EYE , MEL_D) and Oca2 gene within the QTL D. QTL for amino-acid sensitivity (AAsense) and BAC6 within the QTL E. QTL for lens (LensL) in LG3. The plot next to the each linkage map represents the strength of the association of each trait with the markers and the colored bar with the trait label is centered where the LOD score (association with the trait) is maximum. Outlier markers across new and old populations are designated on the map in red color, while those outliers in only one lineage (new or old) are shown in blue color. Thick line on the map represents the region that was used for haplotype phasing. Dotted line on the map labels the region where there is a cluster of the BAC markers. Order of the markers in the overlapping windows was equal to the map position order represented on the map or more detailed in Table 3.4. Markers within the BACs were ordered by the physical positions if the positions between them were known, otherwise the order was arbitrary.
Relationship between different populations groups are shown as percentage of common haplotypes. The haplotype diversity is shown as effective haplotype number (*he*) per each population. Effective haplotype numbers in different populations (*he* plots) are color coded for each population as follows: red represents new cave populations; blue represents old cave populations and grey represents surface populations, as shown in the figure legend of each *he* plot. New cave populations are N1, N2 and N3*, old cave populations are O1, O2O3, O4O6, O8* and surface are SN1 and SN2. In shared haplotypes plots as well as in the plots for significant diversification colors represent following comparisons: red is surface populations *vs.* new caves comparison, blue is surface populations *vs.* old caves comparison and purple is new caves *vs.* old caves comparison as shown in the figure legend. Arrows above sliding window position in each linkage group panel mark the positions of the interest. Significance of the differentiation among cave-surface groups or within groups is shown as a -log10 P-value and is estimated only for the common haplotypes. The dotted line on the graph represents the -log10 P-value of 5%, and all the observations above that limit represent significant diversification across the comparison as estimated by random test and *amova* function in R using 10 000 permutations.

Finally, we also observed a QTL for relative lens size that captured our candidate gene (Fgf8) on LG3 (Figure 3.14.E). This region showed reduction in the number of effective haplotypes and presence of three unique haplotypes, significantly divergent in the new cave population (Figure 3.14.E.ii.,iii.,iv., sliding window position 4, 5 and 6). This is highly suggestive of adaptation specific to the new cave populations. Although we focused only on some of the QTL regions that appear most consistent across the cave populations, we have identified several other markers exhibiting divergence in old, new or both cave populations that might be worth of exploring more in detail. In particular, larger regions of LG7 where no QTL have yet been found have multiple markers that

are differentiated from surface populations (not shown).

## 3.4. DISCUSSION

*Integrated Linkage map-new genomic tool for Astyanax Mexicanus*

To home in on the causative loci, we remapped a subset of the individuals from the original $F_2$ cross that was first mapped using microsatellite markers and used to detect QTL for cave-related phenotypes [77, 78]. The original map was obtained using microsatellite markers from a cross between a surface fish and a Pachón cavefish and the subsequent genotyping and phenotyping of 539 $F_2$ individuals. This map had 29 linkage groups defined by 259 markers and a total linkage map length of 2148cM [77, 78].

Our new linkage map, constructed with only SNP markers greatly improved the map length, reducing it to the 1904cM. The reason for that is probably because of different error rates between two types of markers (SNPs are better than microsatellites) which was also shown in other studies (i.e. [235]). The majority of the SNPs mapped in only a few linkage groups and those linkage groups had a finer distance between the markers (Figure S3.1, Supplementary material). This is probably due to the preferential mapping of the markers that were suitable for genotyping by Sequenom method, as well as markers from BAC clones and candidate genes. The spacing between the markers and distribution of the markers can be improved in the future by using different combinations of restriction enzymes and genotyping more SNPs as shown for other organisms [233, 234].

On the other hand, micro-satellite markers were better spread throughout the genome, thus we used a combination of both types of markers in our linkage map construction in order to maximize genome coverage and provide a finer spaced marker distribution. Both of the maps are used depending on the purpose, but all the references in this study are made to the integrated map, which contained both, SNPs and microsatellites. For example

if we are interested in a QTL region, we use the integrated map and find SNPs in that region. However, some SNPs might not be present in the integrated map so the information about the position of those SNPs relative to ones on the integrated map can be obtained from the SNP only map. In, general the best map improvement was in nine of the linkage groups (LG1`, LG1, LG2, LG3, LG7, LG10, LG12, LG14 and LG25) in terms of marker density and the linkage groups length.

Through usage of the new SNP markers and recalculation of the LOD scores and map positions for the previously measured traits [77], we obtained a great improvement in some regions, especially in terms of better-defined and narrower QTL regions. The LOD profiles obtained only for the linkage groups and traits that were of interest for this study (only those that had outlier SNPs within QTL) show great improvement in LG3 and LG14 for lens and eye QTL. These linkage groups correspond to Protas's LG14 and LG20, respectively [76]. The most obvious improvement was observed in the QTL for eye size in our LG14. With a LOD score of 60, this was our strongest observed QTL, and explained almost 50% of the variance. A QTL for the same trait previously described in LG20 with the LOD score of 30 and ~18% of the variance. The steep increase is centered on m315 SNP marker and the effect strongly decreases towards NYU14, which was previously the center for that QTL with a LOD of ~30 [77].

Another example of the improvement of QTL strength is lens QTL in LG3. This QTL accounted for 13% of the variance in comparison to one in Protas et al. that explained only ~4.2% [77]. This improvement is clearly the result of multiple SNP markers that were placed between 16C and 133B microsatellite makers on our LG3 and narrowed down the QTL to the strong peak of LOD=8 and centered at 103.18 cM (95% CI = 93.8; 112.5). This approach gave us finer candidate regions for some traits that could be explored further in terms of candidate genes. LOD score profiles of the subset

of the LG together with the table for confidence intervals are present in the supplementary material (Table S3.1, Supplementary material).

The integrated map affords us a new tool for *Astyanax* genetics. It is especially valuable for current analyses because we do not yet have a sequenced genome for our species and it is extremely powerful tool to define scaffolds in an eventual genome-sequencing project. One example of this great potential is the genome project for the honey bee (*Apis melifera*), where a linkage map and a genome sequence assembly interactively produced an almost complete organization of the euchromatic genome [236]. Furthermore, BLAST-n analyses of genomic sequences flanking microsatellite markers on the *Astyanax* QTL map against the zebrafish (*Danio reiro*) genome have revealed numerous regions of conserved synteny and it has been extensively used in candidate gene approach by the *Astyanax* community [76]. Similar approaches that combine RAD sequencing with QTL mapping and synteny information could be applied in the other non-model, ecologically interesting species [8, 231, 233].

*SNP discovery, polymorphism and bias corrections in non-model species*

With the capability to generate over a billion bases of DNA sequences per run, next generation sequencing provides one of the best sequencing methods today. Our study, like many other recent studies used the new advances in sequencing technology (paired-end RAD tag sequencing) that allowed us to develop thousands of informative SNPs in the non-sequenced genome of *Astyanax mexicanus* [191, 225-227, 233]. This initial paired-end sequencing of only three individuals permitted us to recover ~300bp contigs around detected SNPs that were used for primer design in more traditional methods of genotyping (genotyping with Sequenom) across larger samples of individuals [237].

The use of SNP markers has widely increased in non-model organisms.

However, their use as a standard tool in population genetics is still challenging, since they are mostly identified in small panels of individuals. As has been shown in this and many other studies, this introduces ascertainment biases [191, 225, 238-243]. This is a potential problem because estimates of population genetic parameters and inferences about demographic processes or scans for the loci with adaptive values can be highly affected by ascertainment bias [134, 240, 243, 244].

In our study we attempted to minimize the potential bias based on the previous knowledge of population history and structure (Chapter 2). It has been established that adjustments for ascertainment bias can be quite effective when such information is available [239]. However, the other panel of markers ("cave SNPs", MAF < 5%) showed the biases towards the polymorphism in old cave populations. Under this scenario, even if the "cave SNPs" markers did not provide any information about the ancestral state of the allele, we were able to use this information based on the variation that was present in multiple independent cave populations. Because there is a low chance that these alleles could arise by mutation at the same nucleotide in the independent populations, they probably came from standing genetic variation in surface populations of either the new or old lineage. However we were not able to detect these polymorphisms in the surface populations; if there, they must be there in low frequency (< 0.05%). In general our power to detect rare variants in ancestral populations was extremely low. Despite our attempts to minimize bias, it might still have affected our data, thus, all our analysis and conclusion were performed strictly based on multiple population comparisons.

*Low level of the polymorphism in the populations*

We observed very low polymorphism in certain cave populations (N1, O4O6) as well as in the SO surface population, regardless of the SNP panel ("cave" or "surface SNPs"). The most intuitive explanation for new cave N1 is that the cave was colonized very recently and the population probably exhibited severe

reductions in population size (small $N_e$) in the recent past, which was already shown based on the microsatellite markers (Chapter 2). The low $N_e$ for the old O4O6 population was also estimated by the microsatellite data that is reflected here in the very low content of polymorphic markers. We believe that this might happen often in natural populations, especially those of endangered species. Comparative analyses of multiple populations, as in our study, will be useful in studies of adaptation in such species [5, 13, 31, 42-44, 46].

Low polymorphism levels in the surface population SO remain difficult to explain. Based on mtDNA analysis this seems to be the surface population we examined that is most closely related to the old cave populations (lineage B in [72]. Therefore it was used in our survey with the expectation that it could serve as a surface sister group to the old cave populations. However, our results do not confirm this expectation. If this is the closest to the old cave populations we would expect to observe a higher level of polymorphism in it than in the old cave populations, as was the case for the new surface to - new caves comparisons. However, that was not the case. This population might belong to an *Astyanax* sister species, and the reduced polymorphism detected could be due to the ascertainment biases in the SNPs collections. Clearly, the status and relationships of this population should be further explored since it is interesting that this is the only surface population of the old mtDNA that was detected in Sierra de El Abra area [72].

*Outlier loci*

Advances in population genomics also made genome-wide screens for adaptive loci feasible in the species without the sequence of the physical genome [238, 245-248]. For example, six natural populations of white spruce (*Picea glauca*) moderately differentiated for several quantitative characters were genotyped for 534 SNPs within 345 candidate genes. Estimation of differentiation in SNP frequencies among populations ($F_{ST}$) revealed 20 SNPs

potentially involved in adaptation [249]. Another study in Atlantic cod also identified multiple loci using different methods for outlier detection, suggesting parallel adaptive evolution [238]. Wright-Fisher fixation indexes, especially the estimator of the population differentiation ($F_{ST}$), have been found to be very robust in many demographic scenarios and allow for neutral loci to be distinguished from the loci with atypical behavior (outlier loci) [128, 139, 154, 156, 157, 163, 168, 229]. Furthermore, the application of these statistics to hierarchically structured populations advanced detection of potentially selected loci in natural systems with more complex population structure [154, 164, 165]. These methods based on population differentiation tests are thus worthwhile for the first insights into adaptive loci in natural populations (i.e [169]).

Our study investigated the genetic framework of adaptation to a cave environment by means of a genome scan for the loci with outlier behavior based on 518 SNP markers. We looked for loci diverging from neutral expectations when comparing populations between the old and new lineages, with the multiple repeatable events of adaptation within each lineage group. All together, 80 outlier loci were identified as potentially involved in adaptation to the cave environment because they were detected in three or more independent populations. This result confirmed the power of the outlier method to reveal signatures of potentially selected loci when multiple comparisons are available and when the association between demography and selection may be complex and/or cryptic like in our system.

Other population genetic causes besides natural selection, such as disassortative mating or Wahlund effect, can also cause outlier behavior [154, 156, 157, 164, 167]. Thus we also explored our data for these sources of deviations from the equilibrium. For example, N3* and O8* populations showed admixture between cave and surface individuals in our microsatellite study (Chapter 2) and we did not expect loci to be in HWE proportions within either of the two populations. Consistent with this expectation, we identified many loci

out of HW in these populations (N3* and O8*) as well as many loci with $F_{IS} > 0$. Thus, there might be some influence of the admixture on the estimation of the outliers, due to Wahlund effect.

We detected multiple loci as outliers in independently derived cave populations. It was very obvious as well that the same loci were observed within each lineage (Table S3.2, Supplementary material). While these coincidences might arise through chance events, the most parsimonious explanation for these observations is that the same genomic regions were involved in parallel adaptations within each lineage. Convergent loci across different lineages were also present and they suggest that despite the different evolutionary history of the two lineages some loci are more liable to change than others. For example, similar observation was also present in the recent butterfly studies where different species of butterflies use the same loci to control color switch [44, 46].

*Linkage disequilibrium and haplotype structure*

QTL mapping has played an important role in discovering many traits and loci in the *Astyanax* genome [77]. However, co-segregating region in the genome can only be determined between closely linked marker and the causative locus. Typically, the number of informative meiosis in crosses is limiting thus, the QTLs have relatively large intervals (Figure 3.14.) [8, 32, 106, 112, 119]. In species with long generation times like *Astyanax* (~ 6 months) the classical cross approach to shorten the QTL region would be almost impossible (reviewed in [8, 106]).

Linkage disequilibrium analysis (LD) is an alternative and powerful method to narrow down the mapping interval because it exploits the segregation of variants in natural populations. Historical recombination stored in the natural populations represents more meiosis and therefore can yield higher resolution maps and can give us insights into the regions that might be

of adaptive value [57, 172, 173, 176]. Despite the wide usage of the $r^2$ measure of the LD across different taxa there is only limited information that we were able to extract on the extent of LD in *Astyanax* genome (Figure 3.12). Because of the paucity of high variability markers (MAF > 5%) in cavefish populations, only a small number of markers were available to test for LD. The distribution of the SNPs per different linkage group was also highly variable, so it was hard to make reasonable comparisons between the LD in different linkage groups. What we can conclude with certainty though, is that the surface populations do not show patches of LD regions, while in the cave populations the LD zones are extensive. This observation suggests small population sizes ($N_e$) in the caves as well as potential bottlenecks, and it is consistent with the observation from the microsatellite data (Chapter 2). Thus, we have confirmed once again the strong influence of demography on the evolutionary histories of cavefish.

Linkage disequilibrium (LD) analysis is based on the relationship between individual genetic markers and non-phased genotypes and often gives a non-monotonic picture [176, 183]. However, an understanding of linkage structure in the cavefish genome is of direct relevance for identification of genes and mutations affecting the traits of interest. Therefore, we further explored haplotype diversities to better understand linkage structure in the cavefish genome. Our study focused on already detected QTL in which the $F_{ST}$ outlier SNPs had previously been identified and we defined our haplotype blocks within such regions. However, we have observed a poor resolution because the markers were too far apart. Thus, the causative loci might be linked to far away regions and inference of the causative region is hard to establish.

Haplotype sharing between old and new populations was mostly low, while between surface and new cave populations was high. Within each lineage we have also observed extensive haplotype sharing. This difference

between the two comparisons is probably the result of different evolutionary history. Furthermore, this suggests that parallel evolution on the genetic level occurs more between closely related populations [2, 24, 250].

Haplotype analysis also finds evidence for small effective population sizes in cave populations reflected in small numbers of effective haplotypes and high levels of haplotype homozygosity. Those differences can also be a result of the small effective population sizes and thus random fixation of the alleles. However, because of numerous haplotypes occurring repeatedly in several populations we argue for the selection on those haplotypes.

*Integrative approach towards detection of natural selection in Cavefish genome*

This is the first study based on a genome-wide scan that addresses the role of the natural selection and neutral forces (drift, migration) in the convergent phenotypic evolution in natural populations of Mexican blind cavefish. The mechanism of regressive evolution in the cavefish, *Astyanax mexicanus* has been the subject of discussion for a very long time [65, 77, 97, 98, 251-254]. Multiple hypotheses have been proposed to explain the regression of the traits in the cavefish. Most refer to eye or pigmentation degeneration, which are the major traits subjected to the loss in the cavefish [77, 254]. Ultimately, all of the hypotheses reduce to either neutral mutations and genetic drift or natural selection as driving forces [97, 253, 254]. Detection of quantitative trait loci (QTL) involved in the eye degeneration and inferences about the same QTL polarity suggest that eye degeneration was driven by natural selection. In contrast, the genetic basis of pigmentation loss is not consistent in polarity, thus pigmentation regression was probably driven by genetic drift [77].

In this study we identified a subset of SNPs and haplotypes at which allele frequencies showed convergence/parallelism in otherwise genetically

distinct natural populations of new and old origin. Previous *Astyanax* studies were mostly based on candidate gene approaches and identified genes associated with morphological traits (functional mutation/deletion), primarily using O1 and N1 crosses [77, 80]. Notably, the majority of those candidate loci surveyed in our population genetic study did not display specific allele or haplotype frequencies. It could well be that our analysis was too conservative (outlier has to appear in at least three populations) and we did not detect those loci or they simply do not have any significant adaptive value. For example, some of these genes are members of the melanin pigmentation pathway (e.g. Mc1r, Oca2); as mentioned above a previously proposed hypothesis suggests that loci involved in pigmentation are rather a product of the mutation accumulation due to genetic drift or relaxed selection in cavefish populations [77, 253]. An exception in the candidate gene survey was fibroblast growth factor (Fgf8) in Lens QTL in LG3 that showed convergence in multiple cave populations. Based on the QTL contribution to the phenotype and signals of differentiation between cave and surface in multiple natural populations, we can suggest that this region is of high adaptive importance in cave populations, and probably subjected to natural selection.

The combined approach of population genetics and QTL analysis is clearly very powerful in non-model organisms [32, 44, 46, 190, 232]. For example, in whitefish (*Coregonus clupeaformis*) differentiation between dwarf and normal ecotypes at growth-associated QTL was shown to be maintained by directional selection [31, 32]. Also, recent study in sticklebacks showed that multiple peaks of selection detected by population genomic screen fall within the previously detected QTLs [5, 7, 9, 233]. An advantage of integrative study using laboratory crosses and natural populations to identify loci under selection was also confirmed in the recent butterfly studies. Population genomic surveys complement QTL mapping approach, and have highlighted gene regions with parallel divergence between forms of the mimetic species on a much finer

scale [42, 44, 46].

In our study we used haplotypes rather than individual SNPs in order to determine if there is haplotype-QTL association and if specific haplotypes show divergence in the cavefish populations. The advantage of this method is that the level of divergence between the populations for the certain regions is given by the multiple SNPs and typically shows more realistic patterns of divergence than those derived from individual SNPs [173, 177, 182-184]. To further investigate the importance of the QTL in LG3 for candidate genes that could be involved in lens phenotype, we have used the synteny information with the *Zebrafish* genome [80]. We performed a BLAST-N search against the *Zebrafish* genome using the contiguous sequences from which our SNP markers were derived. Since those blasted *Astyanax* contigs were highly conserved in order and sequence with the *Danio rerio* genome we were able to obtain homology with ~ 0.7 Mb long regions on chromosome 13 of *Danio rerio* that contains 17 genes. Boundaries of the region represent our phased markers. The overview of the region is shown in the figure 3.15. and individual gene positions and functions from the *Danio rerio* genome are summarized in the Table 3.5.

| *Marker* | *Gene* | *Ch* | *Position* | *Abbreviation* | *Function* |
|---|---|---|---|---|---|
| m15 | Beta-1,3-glucuronyltransferase 2 | 13 | 28175290-28233176 | B3gat2 | Carbohydrate biosynthetic process, expressed in eyes, lens, eye anterior segment, optic nerve, retina |
| | T-cell leukemia, homeobox 1 | 13 | 28491599-28495901 | Tlx1 | Transcription factor activity, neuron differentiation |
| | Solute carrier family 2, member 15a | 13 | 28640899-28665688 | Slc2a15a | Transmembrane transporter activity |
| m703 m702 | Fibroblast growth factor 8 a | 13 | 28719021-28725587 | Fgf8a | Initiation and differentiation of neural retina and lens |
| m17 | LIM-domain binding factor 1a | 13 | 28955703-28969356 | Ldb1a | Transcription cofactor activity |

Table 3.5. Summary of the gene list with their functions and positions from the region on the Chr 13 in *Danio rerio* which shows synteny with cavefish. Marker names are also shown if the gene matched the marker that we have developed in our screening, otherwise only gene names are given.

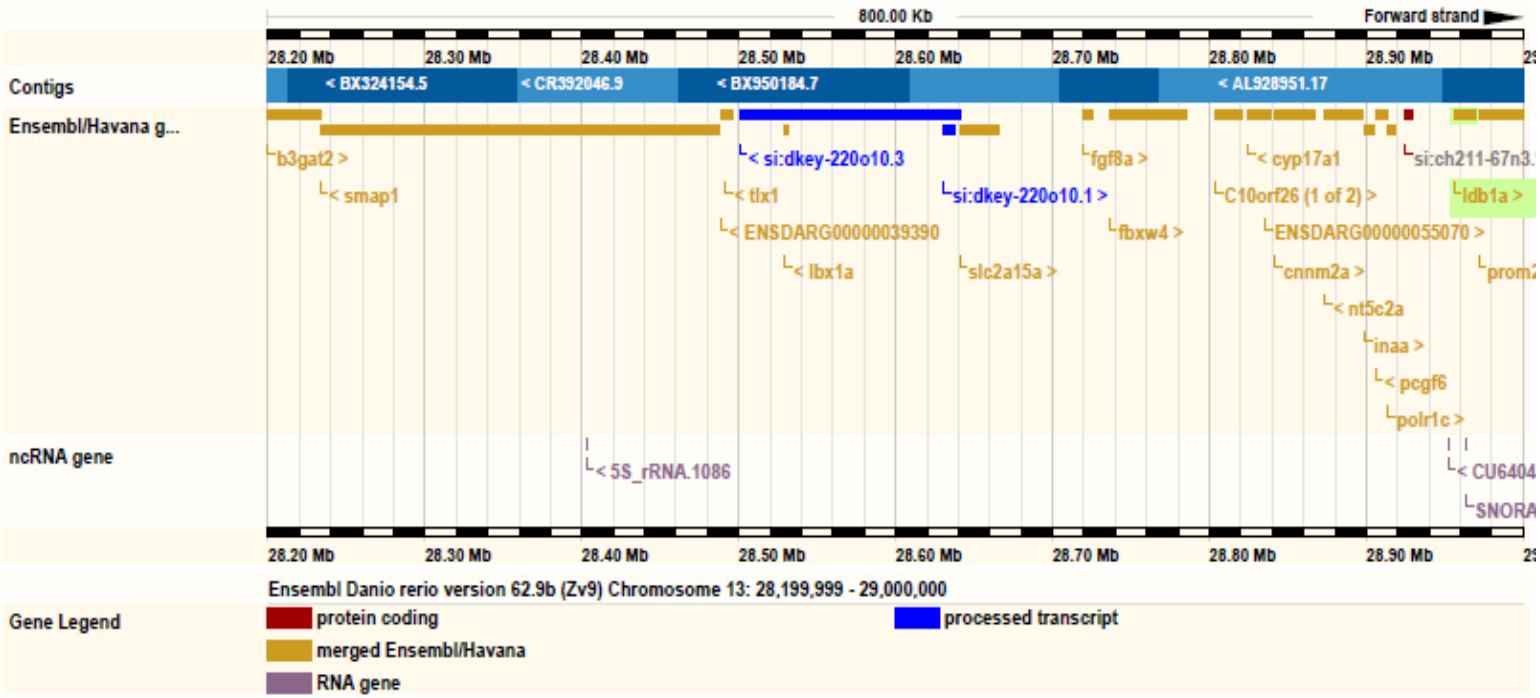



Figure 3.15. Representation of the 800kb of *Danio rerio* region of ZV8 assembly on the chromosome 13 homologous with the QTL region in LG3. Blue rectangular represent processed transcripts, red rectangular represent protein coding genes, yellow rectangular represent merged ensembles, purple rectangular represents RNA genes as shown in the figure legend.

Some of the genes in this region have known involvement in lens development in *Danio rerio* (b3gat2, slc2a15a); the region also contains one of the candidate gene (Fgf8) that was used in our study. Recent studies in cavefish described signaling modifications of Fgf8 and suggest its impact on eye and lens. Furthermore, this study suggests the potential pleiotropic effect associated with this signaling modification that would lead to the evolution of several morphological traits [255].

We have observed the decline in diversity within this QTL region and there are additional evidences of the potential candidate genes in this region. However, due to the high linkage that was observed in the cavefish and because our mapping is not fine enough it is impossible to conclude which region is the main cause of this QTL. It could well be that the observed region is just in strong linkage with some other "causative region". Thus, the targets of adaptation are not clear and the information about the linkage decay is necessary to narrow down the QTL region, which is extremely hard in the populations with the small effective population sizes and low variability [123]. However, this approach might be very useful if there are same adaptive haplotypes with the different lengths in the different populations [182]. By comparing those haplotypes, one could obtain a smaller region with the respective genes.

We have identified significant haplotype divergence also in QTLs for amino-acid sensitivity and overlapping QTL for eye and number of melanophores in LG12 and LG16. Our study supports the hypothesis that the similar adaptive phenotypic changes in different populations can also arise through conserved genetic basis in distantly related lineages (LG12 and LG16 QTL region in new and old cave populations). Thus, we propose those regions to be under strong selection in cave populations in two independent lineages (QTL in LG12 and LG16) as well as within each lineage (QTL in LG3, QTL in BAC1 LG2), suggesting both convergent and parallel evolution in the cave.

The repeated usage of the same genes across distantly related populations and taxonomic groups is already known for some genes in vertebrates (i.e. Mc1r) [13, 113, 115] as well as usage of same loci in distantly related mimetic butterflies [44, 46]. This might suggest that only some regions can evolve to generate particular phenotypic variants for adaptation in multiple species and thus evolutionarily response could be predictable to some extent [256].

Based on the small number of QTLs and markers used in this study we observed higher similarity within each lineage than between the lineages. This suggest that close relatives will overlap more in the details of their adaptive solutions and that forces such as functional constraints, epistasis, and pleiotropy may play an important role in shaping the outcomes of adaptive evolution [2, 24, 250]. That is further shown by the complementary crosses in cavefish [75] as well as for example in sticklebacks [7, 8].

What about the sources of variation through which this adaptations happen? The above-detected haplotypes (in the QTL of LG12 and in LG16) were not detected in the surface populations, while the one in QTL in LG3 was also present in surface population. We reject the hypothesis that these adaptations are due to the new mutations in the cave populations because these haplotypes are shared in the multiple cave populations rather than present as private alleles in the individual populations. We propose instead that the haplotypes selected in the cave are present in very low frequencies in the surface populations, such that we could not detect it in our sample. Therefore, those adaptations are mostly the result of the action of natural selection on standing genetic variation. Very similar phenomena are also observed in multiple freshwater and oceanic stickleback populations [5, 59]. The fact that we have identified only a few QTLs that are affected by natural selection does not necessary mean that other QTLs are not selected for. They simply may not have been identified in our QTL analysis. Our map is relatively sparse and we have also used limited numbers of samples. Furthermore our selection of the

outlier loci was very conservative.

Why did we not find more adaptive haplotypes within the QTL? The QTL studies revealed multiple QTLs for the given phenotypes and they also mostly explained low portion of the phenotypic variance (~10% for the individual QTL). Thus, one of the reasons for that could be explained by the fact that QTLs that we have identified are clearly polygenic so selection might be “diluted” over many loci. In consequence, many of the QTLs behave neutrally, which was also previously described in other systems [31, 32, 257]. Also, QTLs that were identified in laboratory crosses came from the variation that was created by two informative parents which is not necessarily relevant to patterns of selection in natural populations.

Multiple QTL studies have demonstrated the highly polygenic nature of complex phenotypes [104, 159, 258-260]. Furthermore, human association studies (i.e. height) also suggest simultaneous adaptation on multiple loci [261]. This evolutionary mechanism is consistent with short-term adaptation, which is only possible when allele frequencies change at multiple loci at the same time, and strongly supports the concept of adaptation from standing genetic variation [55, 260]. How does adaptation to the cave environment fit into this observation? We have observed that evolutionary history plays an important role in the adaptation to new environment in the cavefish. Thus, cavefish ancestor already harbors significant portion of the variation that could allow for rapid adaptation.

*Conclusion*

In our study we have identified multiple examples of highly divergent loci in different cave populations, ether new, old or both. Some of those loci have been detected as well in the admixed cave populations. All of the methods show that there is divergence of the multiple cave populations on the certain loci, ether by haplotype or by individual SNP. Single and multiple SNP

measures show that there is widespread allele frequency divergence in cave populations but maintenance of diversity within surface populations. Our earlier description of the migration patterns suggests a big migration rate towards cave environment, which is also reflected in the higher diversity of the haplotypes in those populations. Nevertheless, at some loci, local selection seems to be strong enough to overcome the homogenizing effect of gene flow, as shown by detected outliers in those populations. In other cases evidence of potential migrations are low, which also correlates with the lower diversity in those cave populations (N1, O1, N2, and O4O6). This is especially evident in new populations that have probably experienced serious reduction in the population size more recently than the old one. Furthermore, besides the fact that those loci are shared between independently derived populations, they are also identified as QTL loci in laboratory crosses thus implicating that these regions have biological importance.

Even though, none of the specific models that could account for migration, isolation or bottleneck were implemented in this study and it is possible that some complex demographic models could explain our observations, the possibility of divergent selection playing an important role in the cave adaptation should be considered. Thus, the cavefish system is very informative about genetic of adaption in natural populations, but at the same time it is a system with the very complex demographic history and population structure where finding adaptive significance of the loci is a challenging task. However, using the highly repeatable system of morphological change in nature had greatly increased our power towards this effort.

In summary, the extent of (useful) LD measured ether by $r^2$ or haplotype diversity in cave populations could not be adequately estimated, due to high allele frequencies that led to overall high LD. If we assume that the results represented here represent the whole genome, association study would not require genotyping of many SNPs to detect the region of interest that differs

between cave and surface. However, because the inverse relationship between LD and genetic distance is not clear it would probably be hard to use LD for fine-mapping whole genome association study very challenging [183].

## 3.5. MATERIALS AND METHODS

### DNA sampling

*$F_2$ cross samples*

We used DNA samples from a previously described $F_2$ mapping progeny obtained by crossing two full sibs $F_1$ individuals derived from a cross between a wild caught surface fish (Rió Valles, San Luis Potosi) and an individual from the Pachón cave population [78].

*Wild specimens*

We have used previously described populations for which the demographic history and populations structure was determinate. Briefly, populations of cave adapted *Astyanax* populations in NE Mexico are derived from two separate stocks, “old” and “new”. “New” stock consist of all the surface populations plus the cave populations of the Micos area and the Sierra de Guatemala; “Old” stock consists of all the El Abra cave populations from Pachón in the North to Chica in the south (Chapter 2). The names of the populations are derived accordingly with N and O representing “new“ and “old” populations, respectively. Naturally caught fish specimens used in the study were collected during a field trip in March 2008 and preserved in 70% ethanol in the field. Subsequent DNA extraction was done by standardized methods described elsewhere [78]. We collected samples from multiple geographical locations: from caves of the El Abra: O1 (N=32), O2 (N=10), O3 (N=12), O4 (N=12), O6 Curva (N=17), O8* (N=32); from caves of the Sierra de Guatemala: N1 (N=21), N2 (N=26); From Rio Subterraneo cave in the Micos area: N3* (N=25); and from surface localities on the eastern face of the Sierra de el Abra: SN1: SN1a

(N=8), SN1b (N=10), SN1c (N=7), SN1d (N=20), to the South and west of the El Abra, SN2: RSV (N=25) and from the Rascon region to the west SO (N=24). The distribution of the populations and their geographical positions are shown in the map (Figure 2.1) in Chapter 2.

**SNP discovery using RAD tag sequencing method**

The random fragments produced by the RAD tag method [191, 225, 233, 234] were sequenced for the three *Astyanax* $F_1$ parents (cave x surface cross) of the linkage mapping population in order to identify SNPs informative for linkage map construction in the $F_2$ mapping progeny. DNA from each of the three $F_1$ hybrids was digested with high fidelity SbfI (New England Bio labs) and RAD tag libraries were created as in Baird et al. 2008 [233]. The multiplexed *Astyanax* library was placed on a single channel of the Illumina GAII system using 2 x 36bp sequencing chemistry (Paired End Sequencing). Paired-end contigs for each *Astyanax* individual were assembled using Floragenex internal tools. Each RAD paired end sequence contig was compared between the two F1 samples to identify polymorphism. The contigs with the SNP that were heterozygous in both F1 individuals were further tested for genotyping.

*SNP discovery by re-sequencing in BAC fragments and candidate genes*

In order to produce polymorphic markers in longer contiguous regions we produced additional sequences by short insert (~300bp) paired end library construction, IIlumina paired-end sequencing of *Astyanax* bacterial artificial chromosomes (BACs), and *de novo* sequence assembly [225, 226, 262]. The BAC library was commercially available and contains 58,752 clones with the genome coverage of 5X. The genomic DNA fragments range from 45 to 195 Kb and average approximately 105kb [263]. Four BAC clones were sequenced (BAC1, BAC10, BAC6, BACGH); three clones were randomly picked from the library and the fourth was chosen because it had the potential candidate gene, growth hormone (Gh). A standard BAC library screening method was used to

extract the Gh clone [263]. In addition, four other clones, BAC 3, BAC2, BAC7 and BAC 24, were used in trial Solexa Illummina runs, but only small contiguous regions were produced because the reads were only 35bp. Nonetheless, these regions were also considered and a few SNP markers were developed for typing. BAC DNA was purified by Cesium Chloride (*CsCl*) density gradient centrifugation and sequenced by Illumina paired-end sequencing method. Junction sites between BAC backbone and the genomic inserts were used to identify the two ends of the BACs, and BAC vector sequence (~8kb) was removed from the data. The contigs from the Illumina sequencing were assembled *de novo* by Floragenex using internally developed Perl scripts.

In addition to the BAC sequences we obtained flanking sequences around the candidate genes using the Genome Walker kit (Clontech) and gene-specific primers designed using the online software program Primer3 (frodo.wi.mit.edu). Regions inside and around the candidate genes Mc1r, Oca2, Fgf8, Mc1r, and αA-crystalin, which are known to be involved in certain phenotypic changes in cave populations, were surveyed [78, 80, 83, 264]. The SNPs were developed by sequencing the candidate genes and BAC fragments in a panel of 24 individuals from surface and 12 from cave populations. We did direct sequencing of PCR products on an ABI337 automated DNA sequencer (Applied Biosystems) and polymorphic loci were detected by alignment of sequences in BioEdit 7.0.9 [265] and observing the chromatograms by eye. All the SNP polymorphisms for BAC fragments and candidate gene regions were chosen such that the SNP was present in at least 3 surface individuals out of 24 sequenced (~10% MAF) (Figure 3.7).

*SNP genotyping*

We used the MassARRAY mass spectrometry platform from Sequenom (San Diego, USA) to genotype 276 $F_2$ individuals for 675 SNPs and 272 samples

from natural populations for 745 SNPs. PCR-primers and extension-primers were designed using the software MassARRAY Designer (Sequenom) based on following criteria: SNPs must contain 60 bp of flanking genome sequence on each side of the polymorphism**,** no other polymorphism should be present in the 120 base pair sequence landscape, highly repetitive elements (i.e. Alu repeats) or nucleotide based repeats should be avoided during primer design [237]. Assays were multiplexed up to 40 SNPs per reaction. All SNP genotyping was performed according to the standard Sequenom iPLEX protocol. For allele separation, the Sequenom MassARRAY READER instrument (Bruker) was used. Genotypes were assigned based on the presence of mass peaks by the MassARRAY TYPER 4.0 software (Sequenom) [237]. Results were manually inspected, using the MassARRAY Typer Analyzer v4.0 software (Sequenom) (Figure 3.1).

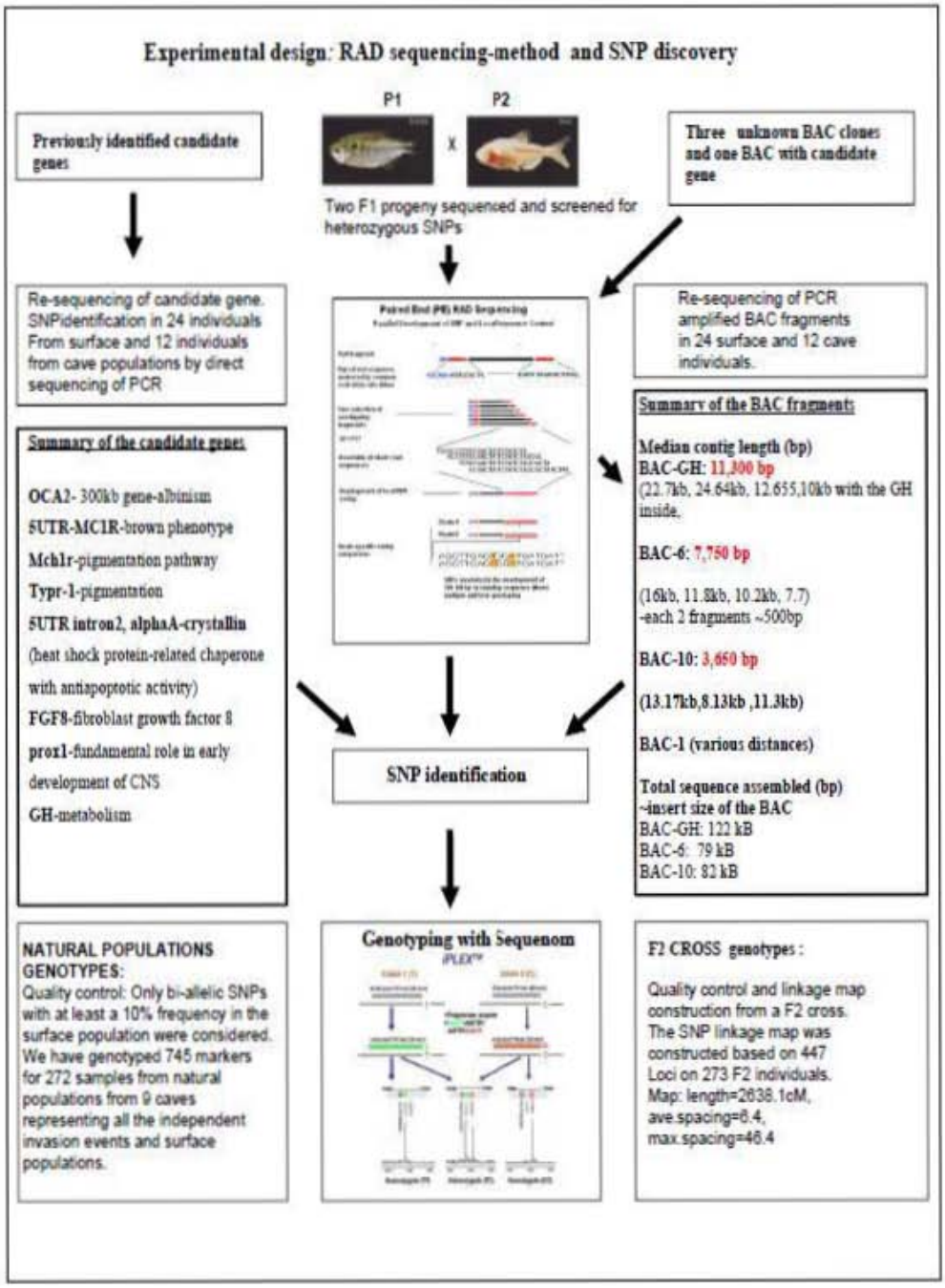



Figure 3.16. Summary of the methods used in SNP discovery (See description in the text).

**Data analysis**

*Quality control*

Quality control was performed on all the markers and 80% of missing data and singleton SNPs were discarded. Individuals with more than 50% of the missing genotypes were discarded from the further analysis. Additional quality control was performed on the genotypes of the surface populations in order to divide the markers in two groups: 1) Minor allele frequency (MAF< 5%) in surface populations and 2) Minor allele frequency (MAF>5%) in surface populations.

*Linkage map construction*

An integrated linkage map for 270 individuals based on 451 SNP and 259 microsatellite markers was established. Data was checked for the segregation patterns and genotypic phase based on the parental genotype. Since the progenitors of the mapping progeny were not inbred but were collected from outbred populations, numerous SNP loci were heterozygous in both parents of the $P_1$ generation. Such loci were not used in map construction because the population origins of the alternative alleles were ambiguous. We used the following criteria to integrate microsatellite and SNP markers into the one linkage map. First, the SNP-only map was constructed and then microsatellite markers were added. The genetic map for SNPs was assessed using JoinMap®4 [266]. The overall approach followed for map generation is described in [77, 78]. Briefly, we performed quality control on the loci that showed abnormal segregation as determined using a chi-square goodness of fit test, and removed those where $P<0.005$. Groups of linked markers were identified using a LOD cut-off value of 5.0 or greater and ordered within linkage groups using Haldane's mapping function.

The corresponding microsatellite and SNP groups were combined using the JoinMap®4 *merge* function to generate a consensus map such that JoinMap®4 was allowed to force additional markers with a lower goodness-of-

fit into the map to maximize the information regarding linked markers for the purposes of this study [266]. We also compared the SNP-only map with maps that contained SNPs and microsatellites, in order to test whether the inclusion of microsatellites caused map length inflation and change in the order.

Re-evaluation of the phenotypic association with the map regions was repeated as in already published data [77] in order to establish the regions of interest related to the quantitative trait. We have recalculated the QTLs for ten traits: relative eye size (RelEye), melanophore numbers (MelLATNEW), dorsal melanophors numbers (MEL_D), melanophore number in eye area (Eyemel), body condition rate (COND), sensitivity to dissolved amino acids (AAsens), rate of weight loss (Wtloss), body length (LEN), ribs number (ribs) and estimated daily growth rate (Grlen) [77]. Detailed description of each trait is given in Table 3.1 A. MultiQTL was used to search for single QTL for the traits listed above in order to determine the LOD scores and proportion of the phenotypic trait variance (PVE-total variance, PEVad-additive variance). PEV and PEVad refer to the proportions of phenotypic trait variance in the mapping progeny ($F_2$) that are explained by a QTL. PEV refers to total trait variance; PEVad refers to the proportion of additive variance explained by the QTL. Non-additive portion of the variance contained in PEV and removed from PEVad refers to interactions among QTLs [267].

First, we identified linkage groups with significant or suggestive trait associations (P =< 0.10) using simple interval mapping (SIM) [107]. Next, all identified linkage groups were used than as a starting set for multiple intervals mapping (MIM) using the *MIM* functions of MultiQTL for each trait [267]. MIM estimates the effects of all detected QTL on each other and uses iteration to estimate the significance of QTL. Using this method background variation is minimized and optimizes estimation of QTL parameters is optimized [268]. Many of the linkage groups in the starting sets had no significant QTL, and these were eliminated from further analysis. Repeating the MIM analysis

generally led to an improvement in significance and precision of the remaining QTL, although some occasionally lost significance. In this case those QTLs were eliminated and the procedure was repeated until a stable set was obtained. The final rounds of MIM analysis for each trait with parameters set at the highest stringency confirmed that the estimates of QTL were consistent. Permutation was used to asses significances of all QTL and confidence intervals on their positions were determined by bootstrap analyses. Significance threshold was set at P = 0.05 for individual QTL, with a genome-wide false detection rate of 10% (FDR = 0.10). These methods detect at most one QTL per trait per linkage group.

*Genetic diversity*

Measures of genetic diversity were estimated for each locus and each population by calculating the percentage of polymorphic SNPs (PO), the mean number of alleles per SNP (A), and the observed ($H_o$) and unbiased expected heterozygosities defined as $H_e = (n / (n - 1)) * 2pq$ where n represents the sample size and p and q frequencies of each allele [269, 270]. Allelic richness and private allelic richness, a measure of the number of alleles independent of sample size and specific for each population was calculated for all populations using HP-Rare [211]. Allelic richness measure is largely influenced by the sample size (e.g. large samples are expected to have more alleles) thus allelic richness and private allelic richness were corrected (rarefied) based on the smallest sample size which was n=14. We also calculated the within-population fixation index $F_{IS} = 1 - (H_o / H_e)$, which provides a measure of the deviation of genotype frequencies (i.e. observed heterozygosity) from Hardy–Weinberg proportions [139, 140]. Deviations from Hardy-Weinberg Equilibrium (HWE) were estimated per each marker in each population using a Fisher exact test as described in [271].

Bonferroni correction was applied in all the analysis used multiple

testing method was used [272]. Due to a high number of monomorphic loci in each cave population, many HW test could not be performed. Thus, thus the proportion of the loci out of HW was calculated as a ratio of significant loci and total number of tests performed per each population.

Since the pooling across the populations did not revealed increasing numbers of departures from HWE, some of the samples were pooled together. Pooled populations were also highly similar in their allelic frequencies and had high migration rate between them as described before (Chapter 2). Populations were pooled for the analysis as follows: cave populations: O2 with O3 in O2O3 population; O4 with O6 in O4O6 population and surface populations were combined in three population pools SN1 = S1 + S2 + S3; SN2 = S4 and SO population was maintained separate since its origin was not determined in our microsatellite study. These names are further retained throughout the chapter.

*Ascertainment bias*

The potential of ascertainment bias due to the SNP discovery having been made from only two F1 individuals for the RAD-tag markers, and from the re-sequencing of only 36 individuals for the candidate gene and BAC fragments analysis, was examined by quantifying diversity, *i.e.* percentage of polymorphic SNPs and $H_e$ in all cave populations relative to surface populations. SNPs were pooled in two groups of RAD tag markers, representing those that were high polymorphic (MAF > 5%) and low polymorphic (MAF < 5%) in the surface population. The same groupings method was applied on the SNP markers derived by re-sequencing. The populations were grouped based on their migration and divergence inferences described in Chapter 2. Correction of ascertainment bias was explored using principal component analysis (PCA) as described in the adgenet package of R [210] by removing the SNPs that were (MAF < 5%) in the surface population.

*Outlier loci detection*

In order to detect potentially adaptive regions in the *Astyanax* genome we tested for outlier loci, i.e., those exhibiting higher $F_{ST}$ values for a given level of heterozygosity than expected from neutral variation using the hierarchical island model as implemented in ARELQUIN 3.5 [154, 165, 213]. We carried out the analysis by comparing two surface populations with individual populations of old or new origin. The comparison was based on the previous structuring of the populations determined by microsatellite data. We accounted for the structuring in the surface populations (SN1, SN2) and contrasted its locus diversity with each of the cave population groups separately (N1, N2, N3*, O1, O2O3, O4O6, O8*). The SO surface population was excluded from this analysis since its polymorphic level was so low as to prevent efficient detection of significant outlier island model implemented in loci. The analysis was performed assuming the presence of 20 groups of 100 demes with 100,000 simulated loci under the model of hierarchical structure [154, 165, 213]. Hierarchical island model implemented in ARELQUIN 3.5 allows the usage of nested analysis of variance and thus accounts for replicated population samples within each hierarchically structured population. When there is a population structure in the data set, the hierarchical model performs better and provides less biased test for outliers than finite model [154, 165]. Finite model assumes that all the populations are independent and can produce false positives. Thus these methods are particularly suitable for our system and purpose of the study. ARELQUIN computes average heterozygosities between populations and compares scaled levels of diversity within and between the populations ($F_{ST}$).

$F_{ST}$ p- values were estimated based on the joined distribution between $F_{ST}$ and $H_e$ using a kernel density function. Diversification of the individual loci based on the $F_{ST}$ measures were calculated per each cave-surface pares. $F_{ST}$ values are shown as functions of expected average heterozygosities ($H_e$).

All the markers ("surface "and "cave") were used in ARELQUIN 3.5 simulations and the significant outliers were sorted after analysis and only markers that were repeatedly identified in the populations were considered.

All the detected loci were classified by their origin: as across lineages outliers (in old and new cave populations) and as lineage specific outliers (only in old or only in new cave populations). Furthermore we have classified each locus conditioned on the presence of variation (MAF > 5%) in surface populations (standing genetic variation (specific for "surface SNPs") outliers.

*Linkage disequilibrium*

The strength of association between alleles at two different markers, pairwise Linkage Disequilibrium (LD), was obtained between all the markers within each linkage group defined from the linkage map. Marker positions at which 5 % < MAF, estimated per each population were excluded from LD calculations to prevent artifacts induced by low frequency alleles. Genetic distances obtained from the linkage map were used as a distances between the markers over which the LD was calculated. LD was calculated as $r^2$, which is a normalized composite genotypic disequilibrium ($D_{AB}$) to address the LD when the genotypic phase of the marker is unknown. Composite coefficient is defined as $D_{AB}=P_{AB} + P_{A/B} - 2p_Ap_B$ where $P_{AB}$ is the frequency of gamete AB, $P_{A/B}$ is the joint frequency of alleles A and B at two different gametes, and $p_A$,$p_B$ are the frequencies of alleles A and B at two loci [173, 175-178]. Decay of LD was described by plotting genetic distances between pairs of SNP markers against their $r^{2,}$

*Phasing and Haplotype frequency*

Regions of interest where defined based on the QTL position in the linkage map. Only QTL regions where the SNP outliers were detected were used in the haplotype phasing. The size of the phased region was defined based on LOD

profile defined by QTL mapping. We have also explored haplotypes diversity based on only BAC fragments or the markers identified only in the candidate genes. Genotypes of unrelated individuals from each population were phased into haplotypes using fastPHASE 1.2. [181]. Fast PHASE uses EM algorithm which is based on maximum likelihood approaches to estimate haplotype phase and we used following settings of the algorithm: 20 random starts, 200 sampled haplotypes from the posterior distribution and 10 cross-validation number of clusters. All the genotypes with posterior probabilities lower than 90% were treated as missing data and were not included in further analysis. We have also performed further quality control by removing haplotypes if there was any ambiguous information about the phase at any single marker in the haplotype (i.e. if the phase could not be inferred for one marker in the haplotype was discarded).

All the haplotypes with the frequency < 5% were discarded in the further analysis. Measures of haplotype diversity were based on relative haplotype frequencies, which were calculated as a ratio between counts for each haplotype and total haplotype number per each population.

Effective number of haplotypes was estimated as $he = 1 / \Sigma pi^2$, with *pi* the frequency of haplotype *i* for a total number of *h* haplotypes. The effective number of haplotypes, analogous to the effective number of alleles [273, 274]. We have performed sliding window analysis of the phased data with 2 SNPs and step size of 1 in the QTL region where at least one significant $F_{ST}$ outlier was found. Next, proportion of common haplotypes for all the populations, new cave-surface, surface-old caves, and new caves-old caves combinations was calculated for QTL from overlapping windows analysis.

Differentiation between haplotype frequencies (haplotype $F_{ST}$) was estimated using *amova* function ade4 package in R [185, 275]. Here, we defined population hierarchy and estimated the variance within and among populations for the given groups of haplotypes derived from the sliding window.

The variance among populations is analogous to Wright's fixation index ($F_{ST}$) and the statistical significance of the variance among populations was evaluated by randomizing the haplotypes over all the populations using *randtest* function in *ade4* package [275]. The proportion of permutations giving an $F_{ST}$ equal to or greater than the observed $F_{ST}$ served as an empirical P-value.

### 3.6. ACKNOWLEDGMENTS

Martina Bradic performed all the experiments, analyzed the data and wrote the chapter. Dr. Richard Borowsky helped with QTL analysis, discussion and suggestions. Dr. Henrique Teotónio supervised project design, experimental design and data analysis. Dr. Ivo Chelo kindly provided his R-scripts and gave suggestions for the data analysis.

# CHAPTER 4

## DISCUSSION

In this thesis work we have 1) identified population structure, divergence, migration patterns and effective population size of multiple cavefish populations, 2) resolved the number of the independent origins of cave related phenotypic trait, 3) developed genome-wide SNP markers using next generation sequencing technology, 4) increased the resolution of the *A. mexicanus* linkage map and increased the resolution of individual QTL loci in surface x cave $F_2$ cross using microsatellite and SNP markers and 5) described the patterns of genetic variation and haplotype structure in natural cave populations across some QTL regions that are associated with phenotypic traits.

### 4.1 Establishing relationships: convergence and parallelism in *Astyanax mexicanus*

Detection of natural selection in the wild is a big challenge due to the unique environments and evolutionary histories of natural populations. As already mentioned, strong demographic influence in natural populations greatly complicates interpretation of natural selection [276-279]. It has been recognized for a long time now that maintenance of the same morphological structure in the nature must be strongly influenced by natural selection and represents the best way for testing natural selection in the wild [17, 34, 50, 51, 131, 280, 281].

In this thesis we are bringing into the focus biological replicates of similar morphology in cavefish (*Astyanax mexicanus*) in order to test natural selection in the wild. As a first step towards understanding adaptation to the cave environment we have disentangled origins and relationships of multiple cave populations that descended from surface fish ancestors (Chapter 2). Our

study supports previous observations that used a variety of molecular data and documented two major colonization events that we have defined as “new” and “old” populations. Population genetic data support the inference that “new” cave populations are closely related to the surface populations and also indicate that present day surface populations can be used as surrogates for stock that gave rise to derived new cave populations [72-74]. On the other hand “old” cave populations are more distantly related to surface and “new” populations. These two lineages diverged about 6.7 Mya, based on estimates from previous studies [72]. In addition to that, our results suggest the old stock surface populations independently inhabited at least three distinct cave localities while there are two independent localities inhabited by “new” stock surface populations (Chapter 2).

In comparison with other fish species, the evolutionary history of the cavefish seem to be different in that it offers opportunity to compare adaptive evolution with closely related lineages (within new and old lineage) and between distantly related lineages of the same species (between new and old lineage). For example, comparative studies of repeated evolution of the same trait in stickleback fishes showed that parallel evolution within this species occurred in freshwater environments colonized by marine sticklebacks after widespread melting of glaciers 10,000 to 20,000 years ago [3, 282-285]. These adaptations are more recent than in the cavefish (few thousand years *vs.* 6.7 Mya). Another widely studied example of parallel adaptation is whitefish. In several lakes across Canada dwarf whitefish have evolved in parallel from a normal whitefish ancestor, starting about 12 000 BP [30]. Both of these examples refer to parallel evolution and represent independent, recent adaptations. Our population genetic study support scenarios of both closely and distantly related populations of the same species independently evolving similar phenotypic traits. Thus, we have determined those relationships as parallel and convergent evolution. Because our goal was to make an explicit

connection between evolution at the phenotypic and genotypic levels we made a definition that bridges both phenotype and genotype. We defined here parallel genotypic adaptation as the independent evolution of same loci responsible for the same function within each lineage. Changes at different loci while comparing different lineages resulting in the same phenotype are considered convergent. Establishment of these relationships was further considered and important while performing comparative studies to test selection. This information allow us to ask whether the signatures of natural selection are present on the same or different loci in different lineages and whether we can assign them to specific phenotypes.

### 4.2. Genetic basis of convergence and parallelism testing selection in the wild

Most studies inquiring into the genetic bases of convergent or parallel evolution in the wild focus mostly on a single gene with a known phenotype (i.e. Mc1r in vertebrates, Oca2 in cavefish, Eda in sticklebacks) [4, 78, 80, 113, 115, 223]. Although very informative, these observations on single genes are problematic, because these studies of natural populations are largely done on pre-selected candidate genes. Also, it might well be that candidate gene studies with no mutational parallelism (at the same site or the same gene) are not frequently reported which gives us a biased view on how frequent those occurrences are (reviewed in [53, 286]). In the summary, most of the empirical studies on the single genes (*see Chapter 1, Table 1.1A and 1.1.B*) showed that closely related populations might evolve the same phenotype using different genes. Also, distantly related organisms, even ones in different classes, might do so using the same genes [2, 13, 53]. It is hard to predict from these observations if generically more related organisms would involve similar genetic changes for the same phenotype that evolved independently. Thus, this information might not reflect general genome-wide patterns and interactions that could lead to

convergence and parallelism. Moving towards genome-wide approaches could bring a completely new view on how same phenotype might arise in distantly and closely related populations [5, 31, 44, 46, 53, 287, 288].

In this thesis we extended our observations to multiple loci. We clearly observed a subset of SNPs and haplotypes exhibiting clear signatures of natural selection in genetically distinct natural populations of new and old origin. We observed haplotypes that were repeatedly selected in cave populations of the new lineage and were present in very low frequencies in the surface populations. On the other hand, within an old lineage we have also observed multiple similar changes (*see Chapter 3*). This suggests that parallel genetic changes are correlated with the level of the relationship between the populations, and that evolutionary history is very important factor in the process of adaptation [27, 28, 58, 289-291]. That would further suggest that parallel genetic change within each lineage is most likely the result of natural selection acting on the standing genetic variation. Similarly, genome-wide studies in sticklebacks show that parallel adaptations in multiple freshwater populations are largely due to standing genetic variation from oceanic stickleback populations (Colosimo, Hosemann et al. 2005; Hohenlohe, Bassham et al. 2010). Parallelism was also observed at the transcriptome level, whereas genes were differentially expressed between normal and dwarf whitefish [288]. Multiple studies of adaptation either in laboratory conditions, or in the wild suggest that the ancestral populations already contain the genetic variation necessary to independently evolve similar phenotypes in response to environmental change (i.e [39, 55, 58, 59, 292]). For example, there will be a greater chance for an advantageous allele to become fixed in a population, rather than lost by genetic drift if the allele is present in multiple copies (standing variation) than if it appears as a new mutation [55, 59, 293]. The probability of fixation will increase with the magnitude of the beneficial effect and with increasing effective population size ($N_e$). In both cases; the probability

of fixation is high for standing variation when it is negligible for new mutations [54, 59]. Beneficial alleles of small effect will especially increase fixation probability from standing variation in small populations such as cavefish.

Assuming that a cave environment would require fast adaptation in a newly established population for at least certain traits, one would expect that the fastest response would be provided through the already available variation [54, 159, 260, 294, 295]. If true, then variation in the same loci and even same sites would be the most prevalent basis for parallel phenotypic diversification among closely related populations. The main evidence for that in the cavefish system, besides identification of the same haplotypes in natural populations are previously performed complementation studies in the lab. This experiments suggest that within the lineage, the complementation of eye and sleep phenotype does not happen or happens only to a very small degree, which would suggest that the genetic changes are very similar and probably occur in parallel [75, 94]. Comparative QTL mapping in different lineages also showed that QTLs mapped to different genomic regions, which further supports that evidence [93].

However, we can also not exclude parallel adaptation from new mutations. For example, we cannot exclude the possibilities that these similarities arose by mutation and that mutation was transmitted to the other cave populations by migration. Independent repeated mutations, and their subsequent fixation in moderately sized or small populations, as is the case for most cave populations, would likely not occur among multiple derived, geographically distant populations [27, 28, 48, 58, 289, 291]. Nevertheless, we also have to consider the possibilities that mutations came independently in the same regions in a different population and because of the similar linkage structure in the cave populations they sweep together with the same haplotypes in different populations. Our resolution was not high enough that we could distinguish among these possibilities.

In our study we have also observed haplotypes within a QTL that are different between the populations in the same lineages, non-parallel genetic change (Chapter 3). The reason for that could be complicated relationships between molecular changes and phenotype due to demography or because different allele is favored in different populations. For example, in a recent genome-wide study of a whitefish morphotypes, no significant overall parallelism between elevated rates of genetic divergence was detected [31]. This has been also reported in other studies that looked at the relationship between selected loci in different populations [26, 296]. Given that selection acts on the phenotype level it is possible that alternative evolutionary trajectories will be taken as selection recruits different alleles, but ultimately lead to the same niche space in the adaptive landscape. In that case the "adaptive alleles" can differ within each lineage but the adaptive significance of the alleles remains consistent (i.e. [31]). For example, in beach mice, similar fur coloration evolved independently through alternative mutations [297]. These examples show that the genetic basis of adaptation can be also highly unpredictable and much more genome-wide data that relate genotype to phenotype will be needed.

We have also found instances of the same haplotypes arising in two different lineages. The observation that different species use the same genetic mechanisms in adaptation to similar environment was also recently shown in mimetic butterfly species. Two species of butterflies, *Heliconius melpomene* and *Heliconius erato* exhibit convergent color patterns wherever they co-occur. Mapping of the color switch genes has revealed that similar phenotypic changes map to the same regions in both species. Thus, similarities in colour pattern have probably evolved through changes in orthologous genes in the two species [44, 46]. Another example represents stickleback populations that evolved reduction in pelvic structures when they invade freshwater habitats. Loss of pelvic structures is associated with a change in expression patterns of

the Pitx1 gene and the same gene is affected in populations along the west coast of Canada, Iceland and Alaska [7, 10, 221, 298]. Recently, Shapiro et al. used expression patterns and intergeneric hybridization to show that Pitx1 is also important in loss of pelvic structure in the distantly related ninespine stickleback (*Pungitius pungitius*) [41]. This would probably suggest that all loci are not equally prone to change and mutations relevant for adaptation tend to accumulate in certain loci or even specific positions within some genes more than in the others [22, 41-43, 113, 299]. That would probably also depend on the specific trait and might suggest constraints by gene function and the structure of genetic networks [2, 13, 42, 43, 45, 300]. One of the most important constraints may be avoidance of negative pleiotropic defects through constrains of coding changes, thus reducing the number of genetic paths adaptation may take [10, 41, 56, 286, 301, 302].

In our study, we did not explore many QTLs for each trait and thus it is hard to predict how those parallel and convergent changes are related to the loci of small or large effect on the genome-wide scale. The polygenic nature underlying the complex traits of *Astyanax* is evident from previous and from our QTL studies (i.e. small effect QTL loci for certain traits, like for example eye size). Another study on QTL mapping of eye size in the cave isopod *Asellu saquaticus*, also showed the polygenic nature of eye size [118]. Thus, polygenic nature of the trait might be common in the cave environment for at least eye size. If true, such an adaptation would occur by small allele frequency shifts spread across many loci. The good example of that is adaptation of human height [159, 303]. There are probably hundreds of SNPs that each affects height by just a few millimeters [261]. Strong selection for height could be very effective, as height is extremely heritable trait. However, at the level of individual SNPs, the effect of selection would be rather weak and would show just a small upward pressure in favor of each of these hundreds of SNPs. This points out that, for a highly polygenic trait, a strong adaptive response could

result from modest allele frequency that is present at many loci. This might be hard to detect in the natural populations where weak effects of each locus could be confounded with the complicated demographic patterns. However, the combination of QTL mapping with population genomics in multiple populations might increase the power of these studies in the future [120, 155, 225, 231].

In summary, our research supports the role of evolutionary history in parallelism and convergence. However, there still is a paucity of data available from the other systems to test for a clear pattern. The synthesis of ecological, phylogenetic, experimental, and genomic advances is promising the answers to at least some of these questions.

### 4.3. Importance of pleiotropy in the evolution of cave related traits

Another very interesting question in the evolution of multiple traits is the importance of indirect selection through the pleiotropic effects. Pleiotropy is defined as a single allelic substitution that alters two or more seemingly, unrelated traits. Its effect is mostly antagonistic such that from one side pleiotropy can increase fitness for certain traits with the trade offs in other fitness related traits [103, 220, 304-308]. Pleiotropy has long been considered a potential mechanism to drive regressive evolution in cave animals through indirect selection [70, 84, 252, 309]. The cavefish model is a powerful system to investigate this question because we can identify numerous traits that evolved together during a defined period of time and also study their genetics [77, 79, 93]. It has been proposed that pleiotropy plays an important role in the evolution of some traits in cavefish [77]. A crucial observation on which this prediction is based is that single trait QTLs for multiple unrelated traits are significantly clustered in the genome of *Astyanax*. Protas et al. argue that the strongest argument that supports the importance of pleiotropy in the cave related traits is diverse phenotypic contents of the QTL clusters and counter-intuitive polarity of substitution effects of some of their constituents [77]. For

example, if more teeth and greater sensitivity to dissolved odorants are of advantage in the cave environment, how would one explain the relatively large negative substitution effects for these traits. Thus, they proposed indirect selection through which these traits could come to predominate in the cave environment [77].

In this thesis, we did not specifically model the possibility of QTL for multiple traits, but our QTL analysis suggest that there might be a reason to think that pleiotropy would also be involved in the process of cave adaption as proposed before [77]. For example, we also identified overlapping regions for QTL for traits that do not seem to be functionally related, such as for example, eye size and the length of the body (Chapetr 3). The very similar observation of multiple traits in the same genomic region was also shown recently in cave isopod *Asellus aquaticus* [118]. QTL mapping in this species also showed that eye size and pigmentation map on the same spot in the genome suggesting co-evolution of these traits. Therefore the genetic architecture of eye and pigment loss might be commonly intertwined in cave animals [118].

One of the possibilities to explain above mentioned phenomenon in cave organisms is that there is already pre-existing cluster of the genes or single genetic locus that segregates at the low frequency in the ancestral population [77]. If beneficial for cave environment, this allele could increase in frequency very rapidly and could influence multiple traits. It is, however, not possible to distinguish between multigenic structure and single genetic locus, because our mapping is not fine enough. Similarly, QTL in *Drosophila* have been shown to be caused by linked genes [310] and data from other model organisms suggest that selection has favored the clustering of genes of related functions [311, 312]. Also, similar phenomena of these gene clusters (i.e. “supergene”) have also been suggested in some plants [313] and wing patterns in butterflies [42, 43, 314]. Chromosomal rearrangements or tight linkage that maintains specific combination of the genes offers the possible

route through which these structures could evolve. However, there are still not enough data available to confirm if this phenomenon is widespread.

Another possibility would be that those pleiotropic effects come from new mutations. It has been shown that small populations will mostly evolve through new mutations of large phenotypic effect, because mutations of small effect are normally present in a very low frequency in the population and will be typically lost by drift (i.e. [315]. The input of new mutations in every generation in small populations is very low. Because of that, there are few options for small populations to select for highly beneficial alleles. Thus, they typically select for the mutations far from the optimum, which might lead to pleiotropy (reviewed in [286]).

Based on our observations, it would be premature to argue that the co-segregation of QTL affecting different traits in the same population is the result of pleiotropy. If pleiotropy is the mechanism through which some of these changes happen, we should be able to rescue multiple traits in cavefish by wild-type alleles. However, this would require having a well-established transgenic system, something still unavailable in *Astyanax mexicanus*.

### 4.4. Perspectives

The completion of our primary research objective advanced our knowledge of the loci underlying phenotypic variation in *Astyanax mexicanus.* In our population genomic scan we have identified putative regions of adaptive significance that no previous mapping approach has identified.

In the first place we would like to note that in order for many further interesting questions to be asked and answered in this impressive natural system; one would undoubtedly need a sequenced genome. With the sequenced genome those regions might be further explored in order to find potential candidate genes. An important implication here is that the results of a combination of laboratory crosses and population genetic scans might increase the resolution

of QTL loci.

Here, we did not explore or use the map of the other two available crosses [93] that are not as advanced as the one used here, because of the number of markers and individuals. However, using three parallel crosses with the population genomics and a sequenced cavefish genome will be very powerful model to find genes that might play an important role in cave adaptation. Also, further convergence and parallel genetic mechanism could be addressed in more detail than here. Recently, *Astyanax* came into focus as a model for the evolution of behavioral traits (i.e. sleep and vibration response) [94, 95]. Thus, the new markers and maps designed here will certainly be of use in the genetic crosses for discovering new behavioral traits.

Many studies showed the big importance of differential gene expression between cave and surface fish, especially in the eye development (reviewed in [316]). Availability of the new sequencing technology will probably also allow exploring the importance of parallel and convergent gene expression or allele specific expression in the evolution of the similar morphologies [191, 225, 288, 317].

If the individual genetic components of the quantitative traits can be dissected and lead to the discovery of new genes and pathways affecting the traits it would be crucial as well to develop techniques in *Astyanax* that would allow testing their functionality in *vivo*. Thus, transgenesis, ectopic expression and morpholino transcriptional knockdown would be very helpful tools to perform functional tests of the putative causal variants [8, 318]. Unfortunately, these techniques are, as yet, only very poorly developed in the cavefish system. Hopefully the community will continue to advance these approaches.

Natural systems in which multiple instances of evolution of the same trait are observed are especially powerful to infer mechanisms of natural selection at the genetic level, since in these cases detection of selected loci could rarely be confounded with demographic effects. The main problem

remains that there is little comparative genome-wide data. In future years we need more studies on convergence and parallelism to drive conclusions and try to understand the rules by which the same or different genetic changes are involved in the similar phenotypic changes. Our research only started to ask this question, by using only a small portion of the genome. However, much more effort must follow in order to answer these important questions.

Our study, as well as few other recent studies [5, 31] suggests that it is extremely important to note that inquiries into the genetic mechanisms underpinning parallel and convergent evolution will have to move beyond single gene observations. The majority of phenotypic characters are complex, and one would expect that complicated developmental pathways and multigenic processes would be involved in this changes. Furthermore, we expect that the role of gene interactions and the importance of pleiotropic effects will further shed light on the genetic mechanisms behind adaptation [319]. The answers to these recurring questions are important in elucidating the mechanisms leading to parallel and convergent evolution.

## REFERENCES


1. Simpson GG: **Principles of Animal Taxonomy**. New York: Columbia University Press; 1961.
2. Wood TE, Burke JM, Rieseberg LH: **Parallel genotypic adaptation: when evolution repeats itself**. *Genetica* 2005, **123**(1-2):157-170.
3. Bell MA, Foster SA: **The Evolutionary Biology of the Threespine Stickleback**. Oxford: Oxford University Press. ; 1994.
4. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Jr., Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM: **Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles**. *Science (New York, NY* 2005, **307**(5717):1928-1933.
5. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags**. *PLoS genetics* 2010, **6**(2):e1000862.
6. Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, Kingsley DM: **The genetic architecture of parallel armor plate reduction in threespine sticklebacks**. *PLoS Biol* 2004, **2**(5):E109.
7. Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH: **Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations**. *Proc Natl Acad Sci U S A* 2004, **101**(16):6050-6055.
8. Cresko WA, McGuigan KL, Phillips PC, Postlethwait JH: **Studies of threespine stickleback developmental evolution: progress and promise**. *Genetica* 2007, **129**(1):105-126.
9. Peichel CL, Nereng KS, Ohgi KA, Cole BL, Colosimo PF, Buerkle CA, Schluter D, Kingsley DM: **The genetic architecture of divergence**

**between threespine stickleback species**. *Nature* 2001, **414**(6866):901-905.

10. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM: **Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks**. *Nature* 2004, **428**(6984):717-723.
11. Chen L, DeVries AL, Cheng CH: **Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod**. *Proc Natl Acad Sci U S A* 1997, **94**(8):3817-3822.
12. Zhang J, Kumar S: **Detection of convergent and parallel evolution at the amino acid sequence level**. *Mol Biol Evol* 1997, **14**(5):527-536.
13. Arendt J, Reznick D: **Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation?** *Trends Ecol Evol* 2008, **23**(1):26-32.
14. Leander BS: **Different modes of convergent evolution reflect phylogenetic distances: a reply to Arendt and Reznick**. *Trends Ecol Evol* 2008, **23**(9):481-482; author reply 483-484.
15. Donoghue MJ: **Key innovations, convergence, and success: macroevolutionary lessons from plant phylogeny**. *Paleobiology* 2005, **31**(2):77-93.
16. Haas O, Simpson GG: **Analysis of Some Phylogenetic Terms, with Attempts at Redefinition**. *Geological Society of America Bulletin* 1945, **56**(12):1164-1164.
17. Simpson GG: **The major features of evolution**. New York: Columbia University Press; 1953.
18. Schluter D: **The ecology of adaptive radiations**. Oxford: Oxford University Press; 2000.
19. Andreev D, Kreitman M, Phillips TW, Beeman RW, ffrench-Constant RH: **Multiple origins of cyclodiene insecticide resistance in**

**Tribolium castaneum (Coleoptera: Tenebrionidae)**. *J Mol Evol* 1999, **48**(5):615-624.

20. Wichman HA, Badgett MR, Scott LA, Boulianne CM, Bull JJ: **Different trajectories of parallel evolution during viral adaptation**. *Science* 1999, **285**(5426):422-424.
21. Ffrench-Constant RH, Anthony N, Aronstein K, Rocheleau T, Stilwell G: **Cyclodiene insecticide resistance: from molecular to population genetics**. *Annu Rev Entomol* 2000, **45**:449-466.
22. Sucena E, Delon I, Jones I, Payre F, Stern DL: **Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism**. *Nature* 2003, **424**(6951):935-938.
23. Sugawara T, Terai Y, Imai H, Turner GF, Koblmuller S, Sturmbauer C, Shichida Y, Okada N: **Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi**. *Proc Natl Acad Sci U S A* 2005, **102**(15):5448-5453.
24. Bollback JP, Huelsenbeck JP: **Parallel Genetic Evolution Within and Between Bacteriophage Species of Varying Degrees of Divergence**. *Genetics* 2009, (181):225–234.
25. Cunningham CW, Jeng K, Husti J, Badgett M, Molineux IJ, Hillis DM, Bull JJ: **Parallel molecular evolution of deletions and nonsense mutations in bacteriophage T7**. *Molecular Biology and Evolution* 1997, **14**(1):113-116.
26. Lenski RE, Stanek MT, Cooper TF: **Identification and dynamics of a beneficial mutation in a long-term evolution experiment with Escherichia coli**. *BMC evolutionary biology* 2009, **9**.
27. Bull JJ, Badgett MR, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux IJ: **Exceptional convergent evolution in a virus**. *Genetics* 1997, **147**(4):1497-1507.

28. Crill WD, Wichman HA, Bull JJ: **Evolutionary reversals during viral adaptation to alternating hosts**. *Genetics* 2000, **154**(1):27-37.
29. Cooper TF, Rozen DE, Lenski RE: **Parallel changes in qene expression after 20,000 generations of evolution in Escherichia coli**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(3):1072-1077.
30. Pigeon D, Chouinard A, Bernatchez L: **Multiple modes of speciation involved in the parallel evolution of sympatric morphotypes of lake whitefish (Coregonus clupeaformis, Salmonidae)**. *Evolution* 1997, **51**(1):196-205.
31. Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L: **SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.)**. *Mol Ecol* 2011, **20**(3):545-559.
32. Rogers SM, Bernatchez L: **Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (Coregonus clupeaformis)**. *Mol Ecol* 2005, **14**(2):351-361.
33. Rogers SM, Bernatchez L: **The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (Coregonus clupeaformis)**. *J Evol Biol* 2006, **19**(6):1979-1994.
34. Rogers SM, Bernatchez L: **The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (Coregonus sp. Salmonidae) species pairs**. *Mol Biol Evol* 2007, **24**(6):1423-1438.
35. Rogers SM, Campbell D, Baird SJ, Danzmann RG, Bernatchez L: **Combining the analyses of introgressive hybridisation and linkage**

**mapping to investigate the genetic architecture of population divergence in the lake whitefish (Coregonus clupeaformis, Mitchill)**. *Genetica* 2001, **111**(1-3):25-41.

36. Rogers SM, Gagnon V, Bernatchez L: **Genetically based phenotype-environment association for swimming behavior in lake whitefish ecotypes (Coregonus clupeaformis Mitchill)**. *Evolution* 2002, **56**(11):2322-2329.
37. Rogers SM, Isabel N, Bernatchez L: **Linkage maps of the dwarf and Normal lake whitefish (Coregonus clupeaformis) species complex and their hybrids reveal the genetic architecture of population divergence**. *Genetics* 2007, **175**(1):375-398.
38. Seehausen O, Terai Y, Sasaki T, Takahashi K, Mizoiri S, Sugawara T, Sato T, Watanabe M, Konijnendijk N, Mrosso HDJ *et al*: **Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids**. *Plos Biology* 2006, **4**(12):2244-2251.
39. Barrett RD, Rogers SM, Schluter D: **Natural selection on a major armor gene in threespine stickleback**. *Science (New York, NY* 2008, **322**(5899):255-257.
40. Schluter D, Conte GL: **Genetics and ecological speciation**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**:9955-9962.
41. Shapiro MD, Bell MA, Kingsley DM: **Parallel genetic origins of pelvic reduction in vertebrates**. *Proc Natl Acad Sci U S A* 2006, **103**(37):13753-13758.
42. Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L *et al*: **Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry**. *Nature* 2011.
43. Joron M, Papa R, Beltran M, Chamberlain N, Mavarez J, Baxter S,

Abanto M, Bermingham E, Humphray SJ, Rogers J *et al*: **A conserved supergene locus controls colour pattern diversity in Heliconius butterflies**. *PLoS Biol* 2006, **4**(10):e303.

44. Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A, Beltran M, Perez-Espona S, Chamberlain N, Ferguson L *et al*: **Genomic hotspots for adaptation: the population genetics of Mullerian mimicry in the Heliconius melpomene clade**. *PLoS genetics* 2010, **6**(2):e1000794.
45. Brakefield PM: **Evo-devo and accounting for Darwin's endless forms**. *Philos Trans R Soc Lond B Biol Sci* 2011, **366**(1574):2069-2075.
46. Counterman BA, Araujo-Perez F, Hines HM, Baxter SW, Morrison CM, Lindstrom DP, Papa R, Ferguson L, Joron M, Ffrench-Constant RH *et al*: **Genomic hotspots for adaptation: the population genetics of Mullerian mimicry in Heliconius erato**. *PLoS genetics* 2010, **6**(2):e1000796.
47. Yoon HS, Baum DA: **Transgenic study of parallelism in plant morphological evolution**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(17):6524-6529.
48. Cohan FM, Hoffmann AA: **Uniform Selection as a Diversifying Force in Evolution - Evidence from Drosophila**. *American Naturalist* 1989, **134**(4):613-637.
49. Rundle HD, Whitlock MC: **A genetic interpretation of ecologically dependent isolation**. *Evolution* 2001, **55**(1):198-201.
50. Schluter D, Clifford EA, Nemethy M, McKinnon JS: **Parallel evolution and inheritance of quantitative traits**. *Am Nat* 2004, **163**(6):809-822.
51. Schluter D, Nagel LM: **Parallel Speciation by Natural-Selection**. *American Naturalist* 1995, **146**(2):292-301.
52. Harvey PH, Pagel MD: **The comparative method in evolutionary**

**biology**. Oxford: Oxford University Press; 1991.

53. Elmer KR, Meyer A: **Adaptation in the age of ecological genomics: insights from parallelism and convergence**. *Trends Ecol Evol* 2011, **26**(6):298-306.

54. Hermisson J, Pennings PS: **Soft sweeps: molecular population genetics of adaptation from standing genetic variation**. *Genetics* 2005, **169**(4):2335-2352.

55. Przeworski M, Coop G, Wall JD: **The signature of positive selection on standing genetic variation**. *Evolution* 2005, **59**(11):2312-2323.

56. Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB: **Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population**. *Science (New York, NY* 2009, **326**(5960):1663-1667.

57. Storz JF, Wheat CW: **Integrating Evolutionary and Functional Approaches to Infer Adaptation at Specific Loci**. *Evolution* 2010, **64**(9):2489-2509.

58. Teotonio H, Chelo IM, Bradic M, Rose MR, Long AD: **Experimental evolution reveals natural selection on standing genetic variation**. *Nat Genet* 2009, **41**(2):251-257.

59. Barrett RD, Schluter D: **Adaptation from standing genetic variation**. *Trends Ecol Evol* 2008, **23**(1):38-44.

60. Orr HA, Betancourt AJ: **Haldane's sieve and adaptation from the standing genetic variation**. *Genetics* 2001, **157**(2):875-884.

61. Jones R, Culver DC, Kane TC: **Are Parallel Morphologies of Cave Organisms the Result of Similar Selection Pressures**. *Evolution* 1992, **46**(2):353-365.

62. Jones R, Culver DC: **Evidence for Selection on Sensory Structures in a Cave Population of Gammarus-Minus (Amphipoda)**. *Evolution* 1989, **43**(3):688-693.

63. Avise JC, Selander RK: **Evolutionary Genetics of Cave-Dwelling Fishes of Genus Astyanax**. *Evolution* 1972, **26**(1):1-&.

64. Breder C: **Descriptive ecology of La Cueva Chica, with especial reference to the blind fish Anoptichthys.** *Zoologica* 1942, **27**:7-15.

65. Mitchell RW, Russell WH, Elliott WR: **Mexican eyeless characin fishes, genus Astyanax: environment, distribution, and evolution**. Lubbock: Texas Tech Press; 1977.

66. Wilkens H: **Genetic Interpretation of Regressive Evolutionary Processes - Studies on Hybrid Eyes of 2 Astyanax Cave Populations (Characidae, Pisces)**. *Evolution* 1971, **25**(3):530-&.

67. Wilkens H: **Evolution and Genetics of Epigean and Cave Astyanax-Fasciatus (Characidae, Pisces) - Support for the Neutral Mutation Theory**. *Evolutionary Biology* 1988, **23**:271-367.

68. Christiansen K, Culver D: **Biogeography and the Distribution of Cave Collembola**. *Journal of Biogeography* 1987, **14**(5):459-477.

69. Culver D, Holsinge.Jr, Baroody R: **Toward a Predictive Cave Biogeography - Greenbrier Valley as a Case Study**. *Evolution* 1973, **27**(4):689-695.

70. Borowsky R, Wilkens H: **Mapping a cave fish genome: Polygenic systems and regressive evolution**. *Journal of Heredity* 2002, **93**(1):19-21.

71. Dowling TE, Martasian DP, Jeffery WR: **Evidence for multiple genetic forms with similar eyeless phenotypes in the blind cavefish, Astyanax mexicanus**. *Molecular Biology and Evolution* 2002, **19**(4):446-455.

72. Ornelas-Garcia CP, Dominguez-Dominguez O, Doadrio I: **Evolutionary history of the fish genus Astyanax Baird & Girard (1854) (Actinopterygii, Characidae) in Mesoamerica reveals multiple morphological homoplasies**. *BMC evolutionary biology* 2008, **8**:340.

73. Strecker U, Bernatchez L, Wilkens H: **Genetic divergence between cave and surface populations of Astyanax in Mexico (Characidae, Teleostei)**. *Molecular Ecology* 2003, **12**(3):699-710.
74. Strecker U, Faundez VH, Wilkens H: **Phylogeography of surface and cave Astyanax (Teleostei) from Central and North America based on cytochrome b sequence data**. *Molecular Phylogenetics and Evolution* 2004, **33**(2):469-481.
75. Borowsky R: **Restoring sight in blind cavefish**. *Curr Biol* 2008, **18**(1):R23-24.
76. Gross JB, Protas M, Conrad M, Scheid PE, Vidal O, Jeffery WR, Borowsky R, Tabin CJ: **Synteny and candidate gene prediction using an anchored linkage map of Astyanax mexicanus**. *Proc Natl Acad Sci U S A* 2008, **105**(51):20106-20111.
77. Protas M, Tabansky I, Conrad M, Gross JB, Vidal O, Tabin CJ, Borowsky R: **Multi-trait evolution in a cave fish, Astyanax mexicanus**. *Evol Dev* 2008, **10**(2):196-209.
78. Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI, Borowsky R, Tabin CJ: **Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism**. *Nature Genetics* 2006, **38**(1):107-111.
79. Jeffery WR: **Regressive evolution in Astyanax cavefish**. *Annual review of genetics* 2009, **43**:25-47.
80. Gross JB, Borowsky R, Tabin CJ: **A Novel Role for Mc1r in the Parallel Evolution of Depigmentation in Independent Populations of the Cavefish Astyanax mexicanus**. *PLoS genetics* 2009, **5**(1).
81. Protas M, Conrad M, Gross JB, Tabin C, Borowsky R: **Regressive evolution in the Mexican cave tetra, Astyanax mexicanus**. *Curr Biol* 2007, **17**(5):452-454.
82. Hooven TA, Yamamoto Y, Jeffery WR: **Blind cavefish and heat shock**

**protein chaperones: a novel role for hsp90 alpha in lens apoptosis**. *International Journal of Developmental Biology* 2004, **48**(8-9):731-738.

83. Strickler AG, Yamamoto Y, Jeffery WR: **Early and late changes in Pax6 expression accompany eye degeneration during cavefish development**. *Development Genes and Evolution* 2001, **211**(3):138-144.

84. Yamamoto Y, Stock DW, Jeffery WR: **Hedgehog signalling controls eye degeneration in blind cavefish**. *Nature* 2004, **431**(7010):844-847.

85. Behrens M, Langecker TG, Wilkens H, Schmale H: **Comparative analysis of Pax-6 sequence and expression in the eye development of the blind cave fish Astyanax fasciatus and its epigean conspecific**. *Molecular Biology and Evolution* 1997, **14**(3):299-308.

86. Behrens M, Wilkens H, Schmale H: **Cloning of the alpha A-crystallin genes of a blind cave form and the epigean form of Astyanax fasciatus: a comparative analysis of structure, expression and evolutionary conservation**. *Gene* 1998, **216**(2):319-326.

87. Strickler AG, Byerly MS, Jeffery WR: **Lens gene expression analysis reveals downregulation of the anti-apoptotic chaperone alphaA-crystallin during cavefish eye degeneration**. *Dev Genes Evol* 2007, **217**(11-12):771-782.

88. Jeffery WR: **Adaptive evolution of eye degeneration in the Mexican blind cavefish**. *Journal of Heredity* 2005, **96**(3):185-196.

89. Jeffery WR, Strickler AG, Yamamoto Y: **To see or not to see: Evolution of eye degeneration in Mexican blind cavefish**. *Integrative and Comparative Biology* 2003, **43**(4):531-541.

90. McCauley DW, Hixon E, Jeffery WR: **Evolution of pigment cell**

**regression in the cavefish Astyanax: a late step in melanogenesis**. *Evol Dev* 2004, **6**(4):209-218.

91. Soares D, Yamamoto Y, Strickler AG, Jeffery WR: **The lens has a specific influence on optic nerve and tectum development in the blind cavefish Astyanax**. *Developmental Neuroscience* 2004, **26**(5-6):308-317.
92. Yamamoto Y, Jeffery WR: **Central role for the lens in cave fish eye degeneration**. *Science (New York, NY* 2000, **289**(5479):631-633.
93. Borowsky R: **The Evolutionary genetics of Cave Fishes:Convergence, Adaptation and Pleiotropy**. In: *Biology of Subterranean Fishes.* Edited by Trajano E, Bichuette, M.E., and Kapoor, B.G.,. Enfield: Science Publishers; 2010: 141-168.
94. Duboue´ E, Keene A, Borowsky R: **Evolutionary Convergence on Sleep Loss in Cavefish Populations**. *Curr Biol* 2011, **21**(1-6).
95. Yoshizawa M, Goricki S, Soares D, Jeffery WR: **Evolution of a Behavioral Shift Mediated by Superficial Neuromasts Helps Cavefish Find Food in Darkness**. *Curr Biol* 2010, **20**(18):1631-1636.
96. Kimura M, Ohta T: **Theoretical aspects of population genetics**. Princeton, N.J.,: Princeton University Press; 1971.
97. Wilkens H: **Genes, modules and the evolution of cave fish**. *Heredity* 2010, **105**(5):413-422.
98. Culver DC, Fong DW: **Why All Cave Animals Look Alike**. *Stygologia* 1986, **2**:208-216.
99. Barr TC: **Speciation in Cave Faunas**. *Annual Review of Ecology and Systematics* 1985, **16**:313-337.
100. Orr HA: **Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data**. *Genetics* 1998, **149**(4):2099-2104.
101. Gelderman H: **Investigation on inheritance of quantitative**

**characters in animals by gene markers. I. Methods**. *Theoretical and Applied Genetics* 1975, **46**:300–319.

102. Kearsey MJ: **The principles of QTL analysis (a minimal mathematics approach)**. *J Exp Bot* 1998, **49**(327):1619-1623.
103. Lynch M, Walsh JB: **Genetics and Analysis of Quantitative Traits**. Massachusetts: Sinauer; 1998.
104. Barton NH, Keightley PD: **Understanding quantitative genetic variation**. *Nature Reviews Genetics* 2002, **3**(1):11-21.
105. Mitchell-Olds T, Willis JH, Goldstein DB: **Which evolutionary processes influence natural genetic variation for phenotypic traits?** *Nat Rev Genet* 2007, **8**(11):845-856.
106. Slate J: **QTL mapping in natural populations: progress, caveats and future directions**. *Molec Ecol* 2005, **14**:363–379.
107. Lander ES, Botstein D: **Mapping Mendelian Factors Underlying Quantitative Traits Using Rflp Linkage Maps**. *Genetics* 1989, **121**(1):185-199.
108. Lander ES, Schork NJ: **Genetic dissection of complex traits**. *Science* 1994, **265**(5181):2037-2048.
109. Barton NH, Turelli M: **Evolutionary quantitative genetic:how little do we know?** *Annual review of genetics* 1989, **23**:337–370.
110. Flint J, Mackay TF: **Genetic architecture of quantitative traits in mice, flies, and humans**. *Genome Res* 2009, **19**(5):723-733.
111. Mackay TFC: **The Genetic-Basis of Quantitative Variation - Numbers of Sensory Bristles of Drosophila-Melanogaster as a Model System**. *Trends in Genetics* 1995, **11**(12):464-470.
112. Long AD, Mullaney SL, Mackay TFC, Langley CH: **Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in Drosophila melanogaster**. *Genetics* 1996, **144**(4):1497-

1510.

113. Nachman MW, Hoekstra HE, D'Agostino SL: **The genetic basis of adaptive melanism in pocket mice**. *Proc Natl Acad Sci U S A* 2003, **100**(9):5268-5273.

114. Mitchell-Olds T: **The molecular basis of quantitative genetic variation in natural populations**. *Trends Ecol Evol* 1995, **10**(8):324-328.

115. Hoekstra HE: **Genetics, development and evolution of adaptive pigmentation in vertebrates**. *Heredity* 2006, **97**(3):222-234.

116. Rosenblum EB, Hoekstra HE, Nachman MW: **Adaptive reptile color variation and the evolution of the Mc1r gene**. *Evolution* 2004, **58**(8):1794-1808.

117. Zhu M, Zhao S: **Candidate gene identification approach: progress and challenges**. *Int J Biol Sci* 2007, **3**(7):420-427.

118. Protas ME, Trontelj P, Patel NH: **Eye and pigment loss in the isopod cave crustacean, Asellus aquaticus**. *Integrative and Comparative Biology* 2011, **51**:E240-E240.

119. Stinchcombe JR, Hoekstra HE: **Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits**. *Heredity* 2008, **100**(2):158-170.

120. Nadeau NJ, Jiggins CD: **A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations**. *Trends Genet* 2010, **26**(11):484-492.

121. Schlotterer C: **Hitchhiking mapping--functional genomics from the population genetics perspective**. *Trends Genet* 2003, **19**(1):32-38.

122. Slatkin M: **Gene flow and the geographic structure of natural populations**. *Science* 1987(236):787–792.

123. Charlesworth B: **Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation**.

*Nat Rev Genet* 2009, **10**(3):195-205.

124. Kirkpatrick M, Barton NH: **Evolution of a species' range**. *Am Nat* 1997, **150**(1):1-23.
125. Lenormand T: **Gene flow and the limits to natural selection**. *Trends in Ecology & Evolution* 2002(17):183–189.
126. Lande R: **Neutral theory of quantitative genetic variance in an island model with local extinction and colonization**. *Evolution* 1992, **46**:381–389.
127. Felsenstein J: **Theoretical Population-Genetics of Variable Selection and Migration**. *Annual review of genetics* 1976, **10**:253-280.
128. Holsinger KE, Weir BS: **Genetics in geographically structured populations: defining, estimating and interpreting F(ST)**. *Nat Rev Genet* 2009, **10**(9):639-650.
129. Pearse DE, Arndt AD, Valenzuela N, Miller BA, Cantarelli V, Sites JW, Jr.: **Estimating population structure under nonequilibrium conditions in a conservation context: continent-wide population genetics of the giant Amazon river turtle, Podocnemis expansa (Chelonia; Podocnemididae)**. *Mol Ecol* 2006, **15**(4):985-1006.
130. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**(2):945-959.
131. Nielsen R: **Molecular signatures of natural selection**. *Annu Rev Genet* 2005, **39**:197-218.
132. Pavlidis P, Hutter S, Stephan W: **A population genomic approach to map recent positive selection in model species**. *Mol Ecol* 2008, **17**(16):3585-3598.
133. Oleksyk TK, Smith MW, O'Brien SJ: **Genome-wide scans for footprints of natural selection**. *Philos Trans R Soc Lond B Biol Sci* 2010, **365**(1537):185-205.

134. Nielsen R: **Estimation of population parameters and recombination rates from single nucleotide polymorphisms**. *Genetics* 2000, **154**(2):931-942.
135. Griffiths RC: **Lines of descent in the diffusion approximation of neutral Wright-Fisher models**. *Theor Popul Biol* 1980, **17**(1):37-50.
136. Gillespie JH: **Population Genetics: A Concise Guide**. In. Baltimore, MD: Johns Hopkins University Press; 1998.
137. Wright S: **Evolution and the genetics of populations**. Chicago: University of Chicago Press; 1978.
138. Wright S: **The Genetical Structure of Populations**. *Annals of Eugenics* 1951, **15**(4):323-354.
139. Weir BS, Cockerham CC: **Estimating F-Statistics for the Analysis of Population-Structure**. *Evolution* 1984, **38**(6):1358-1370.
140. Nei M: **F-statistics and analysis of gene diversity in subdivided populations**. *Ann Hum Genet* 1977, **41**(2):225-233.
141. Wright S: **Evolution in Mendelian populations**. *Genetics* 1931, **16**:97–159.
142. Whitlock MC, McCauley DE: **Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1)**. *Heredity* 1999, **82 ( Pt 2)**:117-125.
143. Teshima KM, Coop G, Przeworski M: **How reliable are empirical genomic scans for selective sweeps?** *Genome Research* 2006, **16**(6):702-712.
144. Hudson RR: **Genetic Data Analysis. Methods for Discrete Population Genetic Data. Bruce S. Weir. Sinauer, Sunderland, MA, 1990. xiv, 377 pp., illus. $48; paper, $27**. *Science (New York, NY* 1990, **250**(4980):575.
145. Tavare S: **Line-of-descent and genealogical processes, and their applications in population genetics models**. *Theor Popul Biol* 1984,

**26**(2):119-164.

146. Fisher R: **The genetical theory of natural selection**. Oxford: Clarendon; 1930.
147. Kingman JFC: **The coalescent**. *Stoch Proc Appl* 1982(13):235–248.
148. Berger JO: **Statistical Decision Theory and Bayesian Analysis**. New York,: Springer; 1985.
149. Avise JC, Jones AG, Walker D, DeWoody JA: **Genetic mating systems and reproductive natural histories of fishes: lessons for ecology and evolution**. *Annu Rev Genet* 2002, **36**:19-45.
150. Dewoody YD, Dewoody JA: **On the estimation of genome-wide heterozygosity using molecular markers**. *J Hered* 2005, **96**(2):85-88.
151. O'Reilly PT, Canino MF, Bailey KM, Bentzen P: **Inverse relationship between F and microsatellite polymorphism in the marine fish, walleye pollock (Theragra chalcogramma): implications for resolving weak population structure**. *Mol Ecol* 2004, **13**(7):1799-1814.
152. Gonzalez EG, Beerli P, Zardoya R: **Genetic structuring and migration patterns of Atlantic bigeye tuna, Thunnus obesus (Lowe, 1839)**. *BMC evolutionary biology* 2008, **8**:252.
153. Alleaume-Benharira M, Pen IR, Ronce O: **Geographical patterns of adaptation within a species' range: interactions between drift and gene flow**. *J Evol Biol* 2006, **19**(1):203-215.
154. Beaumont MA, Balding DJ: **Identifying adaptive genetic divergence among populations from genome scans**. *Mol Ecol* 2004, **13**(4):969-980.
155. Hohenlohe PA, Phillips PC, Cresko WA: **Using Population Genomics to Detect Selection in Natural Populations: Key Concepts and Methodological Considerations**. *International Journal of Plant Sciences* 2010, **171**(9):1059-1071.

156. Lewontin RC, Krakauer J: **Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms**. *Genetics* 1973, **74**(1):175-195.
157. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P: **The power and promise of population genomics: from genotyping to genome typing**. *Nat Rev Genet* 2003, **4**(12):981-994.
158. Charlesworth D: **Balancing selection and its effects on sequences in nearby genome regions**. *PLoS genetics* 2006, **2**(4):379-384.
159. Pritchard JK, Pickrell JK, Coop G: **The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation**. *Curr Biol* 2010, **20**(4):R208-R215.
160. Beaumont MA, Zhang W, Balding DJ: **Approximate Bayesian computation in population genetics**. *Genetics* 2002, **162**(4):2025-2035.
161. Excoffier L: **The detection of regions of our genome under selection has increasingly relied on the use of genome scans**. *Hum Genomics* 2005, **2**(3):155-157.
162. Excoffier L, Heckel G: **Computer programs for population genetics data analysis: a survival guide**. *Nat Rev Genet* 2006, **7**(10):745-758.
163. Beaumont MA: **Adaptation and speciation: what can F(st) tell us?** *Trends Ecol Evol* 2005, **20**(8):435-440.
164. Beaumont MA, Nichols RA: **Evaluating Loci for Use in the Genetic Analysis of Population Structure**. *Proc R Soc Lond B* 1996 (263):1619-1626.
165. Excoffier L, Hofer T, Foll M: **Detecting loci under selection in a hierarchically structured population**. *Heredity* 2009, **103**(4):285-298.
166. Foll M, Beaumont MA, Gaggiotti O: **An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study**

**population structure**. *Genetics* 2008, **179**(2):927-939.

167. Hess JE, Matala AP, Narum SR: **Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin**. *Mol Ecol Resour* 2011, **11 Suppl 1**:137-149.
168. Narum SR, Hess JE: **Comparison of F(ST) outlier tests for SNP loci under selection**. *Mol Ecol Resour* 2011, **11 Suppl 1**:184-194.
169. Shimada Y, Shikano T, Merila J: **A High Incidence of Selection on Physiologically Important Genes in the Three-Spined Stickleback, Gasterosteus aculeatus**. *Molecular Biology and Evolution* 2011, **28**(1):181-193.
170. Wachowiak W, Prus-Glowacki W: **Different patterns of genetic structure of relict and isolated populations of endangered peat-bog pine (Pinus uliginosa Neumann)**. *J Appl Genet* 2009, **50**(4):329-339.
171. Flint-Garcia SA, Thornsberry JM, Buckler ESt: **Structure of linkage disequilibrium in plants**. *Annu Rev Plant Biol* 2003, **54**:357-374.
172. Huff CD, Harpending HC, Rogers AR: **Detecting positive selection from genome scans of linkage disequilibrium**. *BMC Genomics* 2010, **11**:8.
173. Nordborg M, Tavare S: **Linkage disequilibrium: what history has to tell us**. *Trends Genet* 2002, **18**(2):83-90.
174. Ohta T: **Linkage Disequilibrium with the Island Model**. *Genetics* 1982, **101**(1):139-155.
175. Ohta T: **Linkage Disequilibrium Due to Random Genetic Drift in Finite Subdivided Populations**. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* 1982, **79**(6):1940-1944.
176. Slatkin M: **Linkage disequilibrium--understanding the evolutionary**

**past and mapping the medical future**. *Nat Rev Genet* 2008, **9**(6):477-485.

177. Visscher PM: **Variation of estimates of SNP and haplotype diversity and linkage disequilibrium in samples from the same population due to experimental and evolutionary sample size**. *Ann Hum Genet* 2007, **71**(Pt 1):119-126.

178. Weir BS: **Inferences About Linkage Disequilibrium**. *Biometrics* 1979, **35**(1):235-254.

179. Lewontin RC, Kojima K: **The evolutionary dynamics of complex polymorphisms**. *Evolution* 1960(14):458–472

180. Lewontin RC: **On measures of gametic disequilibrium**. *Genetics* 1988, **120**(3):849-852.

181. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase**. *Am J Hum Genet* 2006, **78**(4):629-644.

182. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ *et al*: **Detecting recent positive selection in the human genome from haplotype structure**. *Nature* 2002, **419**(6909):832-837.

183. Wall JD, Pritchard JK: **Haplotype blocks and linkage disequilibrium in the human genome**. *Nat Rev Genet* 2003, **4**(8):587-597.

184. Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA: **Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations**. *Am J Hum Genet* 2004, **74**(4):610-622.

185. Excoffier L, Smouse PE, Quattro JM: **Analysis of molecular variance inferred from metric distances among DNA haplotypes:**

**application to human mitochondrial DNA restriction data**. *Genetics* 1992, **131**(2):479-491.

186. Breseghello F, Sorrells ME: **Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars**. *Genetics* 2006, **172**(2):1165-1177.

187. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Maize association population: a high-resolution platform for quantitative trait locus dissection**. *Plant J* 2005, **44**(6):1054-1064.

188. Rafalski A, Morgante M: **Corn and humans: recombination and linkage disequilibrium in two genomes of similar size**. *Trends Genet* 2004, **20**(2):103-111.

189. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC *et al*: **The genetic architecture of maize flowering time**. *Science (New York, NY* 2009, **325**(5941):714-718.

190. Campbell D, Bernatchez L: **Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes**. *Mol Biol Evol* 2004, **21**(5):945-956.

191. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML: **Genome-wide genetic marker discovery and genotyping using next-generation sequencing**. *Nat Rev Genet* 2011, **12**(7):499-510.

192. Manceau M, Domingues VS, Mallarino R, Hoekstra HE: **The developmental role of Agouti in color pattern evolution**. *Science (New York, NY* 2011, **331**(6020):1062-1065.

193. Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J *et al*: **Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer**. *Science (New York, NY* 2010, **327**(5963):302-

305.

194. Nevo E, Beiles A, Spradling T: **Molecular evolution of cytochrome b of subterranean mole rats, Spalax ehrenbergi superspecies, in Israel.** . *Journal of Molecular Evolution* 1999, **49**:215-226.

195. Evanno G, Regnaut S, Goudet J: **Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study**. *Molecular Ecology* 2005, **14**(8):2611-2620.

196. Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibly RM: **Likelihood-based estimation of microsatellite mutation rates**. *Genetics* 2003, **164**(2):781-787.

197. Yue GH, David L, Orban L: **Mutation rate and pattern of microsatellites in common carp (Cyprinus carpio L.)**. *Genetica* 2007, **129**(3):329-331.

198. Beerli P, Palczewski M: **Unified framework to evaluate panmixia and migration direction among multiple sampling locations**. *Genetics* 2010, **185**(1):313-326.

199. Fitzsimmons K: **Tilapia aquaculture in Mexico**. In: *Tilapia Aquaculture in the Americas.* Edited by Costa-Pierce B, Rakocy J, vol. 2. Baton Rouge: The World Aquaculture Society; 2000: 171-183.

200. Panaram K, Borowsky R: **Gene flow and genetic variability in cave and surface populations of the Mexican Tetra, Astyanax mexicanus (Telcostei : Characidae)**. *Copeia* 2005(2):409-416.

201. Espinasa L, Borowsky R: **Eyed cave fish in a karst window**. *Journal of Cave and Karst Studies* 2000, **62**:180-183.

202. Magalhaes IS, Mwaiko S, Seehausen O: **Sympatric colour polymorphisms associated with nonrandom gene flow in cichlid fish of Lake Victoria**. *Mol Ecol* 2010, **19**(16):3285-3300.

203. Nei M: **Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals**. *Genetics* 1978, **89**(3):583-590.

204. Raymond M, Rousset F: **Genepop (Version-1.2) - Population-Genetics Software for Exact Tests and Ecumenicism**. *Journal of Heredity* 1995, **86**(3):248-249.

205. Dieringer D, Schlotterer C: **MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets**. *Molecular Ecology Notes* 2003, **3**(1):167-169.

206. Guo SW, Thompson EA: **Performing the exact test of Hardy-Weinberg proportion for multiple alleles**. *Biometrics* 1992, **48**(2):361-372.

207. Rice WR: **Analyzing Tables of Statistical Tests**. *Evolution* 1989, **43**(1):223-225.

208. Huelsenbeck JP, Andolfatto P: **Inference of population structure under a Dirichlet process model**. *Genetics* 2007, **175**(4):1787-1802.

209. Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE: **Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation**. *Conservation Genetics* 2006, **7**(2):295-302.

210. Jombart T: **adegenet: a R package for the multivariate analysis of genetic markers**. *Bioinformatics (Oxford, England)* 2008, **24**(11):1403-1405.

211. Kalinowski ST: **HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness**. *Molecular Ecology Notes* 2005, **5**(1):187-189.

212. Leberg PL: **Estimating allelic richness: Effects of sample size and bottlenecks**. *Molecular Ecology* 2002, **11**(11):2445-2449.

213. Excoffier L, Lischer HEL: **Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows**. *Molecular Ecology Resources* 2010, **10**(3):564-567.

214. Beerli P: **Comparison of Bayesian and maximum-likelihood inference of population genetic parameters**. *Bioinformatics (Oxford, England)* 2006, **22**(3):341-345.

215. Beerli P, Felsenstein J: **Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(8):4563-4568.

216. Makinen HS, Cano JM, Merila J: **Genetic relationships among marine and freshwater populations of the European three-spined stickleback (Gasterosteus aculeatus) revealed by microsatellites**. *Molecular Ecology* 2006, **15**(6):1519-1534.

217. Rockman MV: **Reverse engineering the genotype-phenotype map with natural genetic variation**. *Nature* 2008, **456**(7223):738-744.

218. Rockman MV, Kruglyak L: **Recombinational Landscape and Population Genomics of Caenorhabditis elegans**. *PLoS genetics* 2009, **5**(3).

219. Rockman MV, Skrovanek SS, Kruglyak L: **Selection at Linked Sites Shapes Heritable Phenotypic Variation in C. elegans**. *Science* 2010, **330**(6002):372-376.

220. Barrett RD, Rogers SM, Schluter D: **Environment specific pleiotropy facilitates divergence at the Ectodysplasin locus in threespine stickleback**. *Evolution* 2009, **63**(11):2831-2837.

221. Bell MA, Orti G: **Pelvic reduction in threespine stickleback from Cook Inlet lakes: geographical distribution and intrapopulation variation**. *Copeia* 1994:314–325.

222. Hoekstra HE, Drumm KE, Nachman MW: **Ecological genetics of adaptive color polymorphism in pocket mice: geographic variation in selected and neutral genes**. *Evolution* 2004, **58**(6):1329-1341.

223. Hoekstra HE, Price T: **Evolution. Parallel evolution is in the genes**. *Science (New York, NY* 2004, **303**(5665):1779-1781.

224. Hoekstra HE, Manceau M, Domingues VS, Linnen CR, Rosenblum EB: **Convergence in pigmentation at multiple levels: mutations, genes and function**. *Philos T R Soc B* 2010, **365**(1552):2439-2450.

225. Stapley J, Reger J, Feulner PG, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J: **Adaptation genomics: the next generation**. *Trends Ecol Evol* 2010, **25**(12):705-712.

226. Hudson ME: **Sequencing breakthroughs for genomic ecology and evolutionary biology**. *Molecular Ecology Resources* 2008, **8**(1):3-17.

227. Mardis ER: **Next-generation DNA sequencing methods**. *Annu Rev Genom Hum G* 2008, **9**:387-402.

228. Foll M, Gaggiotti O: **A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective**. *Genetics* 2008, **180**(2):977-993.

229. Bonin A: **Population genomics: a new generation of genome scans to bridge the gap with functional genomics**. *Molecular Ecology* 2008, **17**(16):3583-3584.

230. Akey JM: **Issues in Detecting Natural Selection in Humans.** *American Journal of Physical Anthropology* 2009:75-75.

231. Phillips PC: **Testing hypotheses regarding the genetics of adaptation**. *Genetica* 2005(123):15–24.

232. Roberge C, Guderley H, Bernatchez L: **Genomewide identification of genes under directional selection: gene transcription Q(ST) scan in diverging Atlantic salmon subpopulations**. *Genetics* 2007, **177**(2):1011-1022.

233. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA: **Rapid SNP discovery and genetic mapping using sequenced RAD markers**. *PLoS One* 2008,

**3**(10):e3376.

234. Davey JW, Blaxter ML: **RADSeq: next-generation population genetics**. *Brief Funct Genomics* 2010, **9**(5-6):416-423.

235. Ball AD, Stapley J, Dawson DA, Birkhead TR, Burke T, Slate J: **A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (Taeniopygia guttata)**. *BMC Genomics* 2010, **11**:218.

236. Solignac M, Zhang L, Mougel F, Li BS, Vautrin D, Monnerot M, Cornuet JM, Worley KC, Weinstock GM, Gibbs RA: **The genome of Apis mellifera: dialog between linkage mapping and sequence assembly**. *Genome Biology* 2007, **8**(3).

237. Bradic M, Costa J, Chelo IM: **Genotyping with Sequenom** In: *Molecular Methods for Evolutionary Genetics.* Edited by Orgogozo V, Rockman M, vol. Vol. 772. New York Humana Press; 2011.

238. Bradbury IR, Hubert S, Higgins B, Bowman S, Paterson IG, Snelgrove PV, Morris CJ, Gregory RS, Hardie DC, Borza T *et al*: **Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, Gadus morhua**. *Mol Ecol Resour* 2011, **11 Suppl 1**:218-225.

239. Schlotterer C, Harr B: **Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in Drosophila melanogaster**. *Mol Ecol* 2002, **11**(5):947-950.

240. Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K: **The discovery of single-nucleotide polymorphisms--and inferences about human demographic history**. *Am J Hum Genet* 2001, **69**(6):1332-1347.

241. Ramirez-Soriano A, Nielsen R: **Correcting estimators of theta and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process**. *Genetics* 2009,

**181**(2):701-710.

242. Rosenblum EB, Novembre J: **Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard**. *J Hered* 2007, **98**(4):331-336.

243. Nielsen R: **Population genetic analysis of ascertained SNP data**. *Hum Genomics* 2004, **1**(3):218-224.

244. Kuhner MK, Beerli P, Yamato J, Felsenstein J: **Usefulness of single nucleotide polymorphism data for estimating population parameters**. *Genetics* 2000, **156**(1):439-447.

245. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, Cariani A, Maes GE, Diopere E, Carvalho GR *et al*: **Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges**. *Mol Ecol Resour* 2011, **11 Suppl 1**:123-136.

246. Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW: **Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms**. *Mol Ecol Resour* 2011, **11 Suppl 1**:1-8.

247. Seeb JE, Pascal CE, Grau ED, Seeb LW, Templin WD, Harkins T, Roberts SB: **Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids**. *Mol Ecol Resour* 2011, **11**(2):335-348.

248. Seeb LW, Templin WD, Sato S, Abe S, Warheit K, Park JY, Seeb JE: **Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon**. *Mol Ecol Resour* 2011, **11 Suppl 1**:195-217.

249. Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J: **Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce**. *Mol Ecol* 2008,

**17**(16):3599-3613.

250. Futuyma DJ: **Evolutionary Biology**. Sunderland, MA: Sinauer Associates.; 1986.

251. Jones R, Culver D, Kane T: **Are parallel morphologies of cave organisms the result of similar selection pressures?** *Evol Int J Org Evol* 1992(46):353–365.

252. Culver DC: **Cave Life: Evolution and Ecology**. CAmbrige: Harvard University Press., Cambridge, Mass.; 1982.

253. Wilkens H: **Variability and loss of functionless traits in cave animals. Reply to Jeffery (2010)**. *Heredity* 2010, **106**(4):707-708.

254. Jeffery WR: **Pleiotropy and eye degeneration in cavefis**. *Heredity* 2010(105):495-496.

255. Pottin K, Hinaux H, Retaux S: **Restoring eye size in Astyanax mexicanus blind cavefish embryos through modulation of the Shh and Fgf8 forebrain organising centres**. *Development* 2011, **138**(12):2467-2476.

256. Orgogozo V, Stern DL: **How different are recently diverged species?: more than 150 phenotypic differences have been reported for the D. melanogaster species subgroup**. *Fly (Austin)* 2009, **3**(2):117.

257. O'Malley KG, Camara MD, Banks MA: **Candidate loci reveal genetic differentiation between temporally divergent migratory runs of Chinook salmon (Oncorhynchus tshawytscha)**. *Mol Ecol* 2007, **16**(23):4930-4941.

258. Hill WG, Caballero A: **Artificial selection experiments**. *Annu Rev Ecol Sys* 1992(23):287–310.

259. Grant PR, Grant BR: **Predicting microevolutionary responses to directional selection on heritable variation**. *Evolution* 1995(49):241–251.

260. Turelli M, Barton NH: **Genetic and statistical analyses of strong selection on polygenic traits: what, me normal?** *Genetics* 1994, **138**(3):913-941.

261. Chatterjee N, Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ: **Estimation of effect size distribution from genome-wide association studies and implications for future discoveries**. *Nature Genetics* 2010, **42**(7):570-U139.

262. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

263. Di Palma F, Kidd C, Borowsky R, Kocher TD: **Construction of bacterial artificial chromosome libraries for the Lake Malawi cichlid (Metriaclima zebra), and the blind cavefish (Astyanax mexicanus)**. *Zebrafish* 2007, **4**(1):41-47.

264. Retaux S, Pottin K, Alunni A: **Shh and forebrain evolution in the blind cavefish Astyanax mexicanus**. *Biol Cell* 2008, **100**(3):139-147.

265. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT**. *Nucleic Acids Symposium Series* 1999, **41**:95-98.

266. Van Ooijen JW, Voorips RE: **Software for the Calculation of Genetic Linkage Maps** Wageningen, Netherlands: Plant Research International; 2001.

267. Korol AB, Ronin YI, Nevo H, Hayes P: **Multi-interval**
**mapping of correlated trait complexes: simulation analysis and evidence from barley. Heredity 80: 273–284.** 1998.

268. Kao CH, Zeng ZB, Teasdale RD: **Multiple interval mapping for quantitative trait loci**. *Genetics* 1999, **152**(3):1203-1216.

269. Nei M: **Analysis of gene diversity in subdivided populations**. *Proc*

*Natl Acad Sci U S A* 1973, **70**(12):3321-3323.

270. Nei M, Roychoudhury AK: **Sampling variances of heterozygosity and genetic distance**. *Genetics* 1974, **76**(2):379-390.
271. Wigginton JE, Cutler DJ, Abecasis GR: **A note on exact tests of Hardy-Weinberg equilibrium**. *Am J Hum Genet* 2005, **76**(5):887-893.
272. Dunn OJ: **Multiple Comparisons Among Means**. *Journal of the American Statistical Association* 1961(56):52-64.
273. Ewens WJ: **The Maintenance of Alleles by Mutation**. *Genetics* 1964, **50**:891-898.
274. Hartl DL, Clark AG: **Principles of population genetics (3rd ed)**. Sunderland, MA Sinauer Associates; 1997.
275. Dray S, Dufour AB: **The ade4 Package: Implementing the Duality Diagram for Ecologists**. *Journal of Statistical Software* 2007, **22**(4):1-20.
276. Gillespie JH: **Genetic drift in an infinite population. The pseudohitchhiking model**. *Genetics* 2000, **155**(2):909-919.
277. Hill WG, Robertson A: **The effect of linkage on limits to artificial selection**. *Genet Res* 1966, **8**(3):269-294.
278. Doebley JF, Gaut BS, Smith BD: **The molecular genetics of crop domestication**. *Cell* 2006, **127**(7):1309-1321.
279. Wright SI, Gaut BS: **Molecular population genetics and the search for adaptive evolution in plants**. *Mol Biol Evol* 2005, **22**(3):506-519.
280. Rundle HD, Nagel L, Wenrick Boughman J, Schluter D: **Natural selection and parallel speciation in sympatric sticklebacks**. *Science (New York, NY* 2000, **287**(5451):306-308.
281. Rieseberg LH, Widmer A, Arntz AM, Burke JM: **Directional selection is the primary cause of phenotypic diversification**. *Proc Natl Acad Sci* 2002(99):12 242–212 245.
282. O'Reilly P, Reimchen TE, Beech R, Strobeck C: **Mitochondrial DNA in**

**Gasterosteus and Pleistocene glacial refugium on the Queen Charlotte Islands, British Columbia**. *Evolution* 1993, **47**:678–684.

283. Ortı´ G, Bell MA, Reimchen TE, Meyer A: **Global survey of mitochondrial DNA sequences in the threespine stickleback: evidence for recent migrations**. *Evolution 49* 1994(49):608–622.
284. Bell MA: **Lateral plate evolution in the threespine stickleback: getting nowhere fast**. *Genetica* 2001, **112**:445-461.
285. Walker JA, Bell MA: **Net evolutionary trajectories of body shape evolution within a microgeographic radiation of threespine sticklebacks (Gasterosteus aculeatus)**. *J Zool* 2000(252):293–302.
286. Stern DL, Orgogozo V: **Is genetic evolution predictable?** *Science (New York, NY* 2009, **323**(5915):746-751.
287. Michel AP, Sim S, Powell TH, Taylor MS, Nosil P, Feder JL: **Widespread genomic divergence during sympatric speciation**. *Proc Natl Acad Sci U S A* 2010, **107**(21):9724-9729.
288. St-Cyr J, Derome N, Bernatchez L: **The transcriptomics of life-history trade-offs in whitefish species pairs (Coregonus sp.)**. *Molecular Ecology* 2008, **17**(7):1850-1870.
289. Teotonio H, Rose MR: **Variation in the reversibility of evolution**. *Nature* 2000, **408**(6811):463-466.
290. Cohan FM: **Can Uniform Selection Retard Random Genetic-Divergence between Isolated Conspecific Populations**. *Evolution* 1984, **38**(3):495-504.
291. Travisano M, Mongold JA, Bennett AF, Lenski RE: **Experimental Tests of the Roles of Adaptation, Chance, and History in Evolution**. *Science (New York, NY* 1995, **267**(5194):87-90.
292. Lindsey CC: **Experimental study of meristic variation in a population of threespine stickleback, Gasterosteus aculeatus**. *Can J Zool 40* 1962a(40):271–312.

293. Innan H, Kim Y: **Pattern of polymorphism after strong artificial selection in a domestication event**. *Proc Natl Acad Sci U SA* 2004(101):10667–10672.

294. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M *et al*: **Convergent adaptation of human lactase persistence in Africa and Europe**. *Nat Genet* 2007, **39**(1):31-40.

295. Feder JL, Berlocher SH, Roethele JB, Dambroski H, Smith JJ, Perry WL, Gavrilovic V, Filchak KE, Rull J, Aluja M: **Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis**. *Proc Natl Acad Sci U S A* 2003, **100**(18):10314-10319.

296. Nolte AW, Gompert Z, Buerkle CA: **Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations**. *Molecular Ecology* 2009, **18**(12):2615-2627.

297. Steiner CC, Rompler H, Boettger LM, Schoneberg T, Hoekstra HE: **The Genetic Basis of Phenotypic Convergence in Beach Mice: Similar Pigment Patterns but Different Genes**. *Molecular Biology and Evolution* 2009, **26**(1):35-45.

298. Bell MA: **Interacting evolutionary constraints in pelvic reduction of threespine sticklebacks, Gasterosteus aculeatus (Pisces, Gasterosteidae)**. *BiolJ Linn Soc* 1987, **31**:347–382.

299. Theron E, Hawkins K, Bermingham E, Ricklefs RE, Mundy NI: **The molecular basis of an avian plumage polymorphism in the wild: a melanocortin-1-receptor point mutation is perfectly associated with the melanic plumage morph of the bananaquit, Coereba flaveola**. *Curr Biol* 2001, **11**(8):550-557.

300. Gibson G, Wagner G: **Canalization in evolutionary genetics: a stabilizing theory?** *Bioessays* 2000, **22**(4):372-380.

301. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW: **Crystal**

**structure of an ancient protein: evolution by conformational epistasis**. *Science (New York, NY* 2007, **317**(5844):1544-1548.

302. Weinreich DM, Delaney NF, DePristo MA, Hartl DL: **Darwinian evolution can follow only very few mutational paths to fitter proteins**. *Science*, 2006(312):11-114.

303. Pritchard JK, Di Rienzo A: **Adaptation - not by sweeps alone**. *Nature Reviews Genetics* 2010, **11**(10):665-667.

304. Falconer DS, Mackay TFC: **Introduction to quantitative genetics**. Essex, UK: Longman; 1996.

305. Barton NH: **Pleiotropic models of quantitative variation**. *Genetics* 1990, **124**(3):773-782.

306. Griswold CK, Whitlock MC: **The genetics of adaptation: the roles of pleiotropy, stabilizing selection and drift in shaping the distribution of bidirectional fixed mutational effects**. *Genetics* 2003, **165**(4):2181-2192.

307. Otto SP: **Two steps forward, one step back: the pleiotropic effects of favoured alleles**. *Proc Biol Sci* 2004, **271**(1540):705-714.

308. Keightley PD, Hill WG: **Quantitative genetic variability maintained by mutation-stabilizing selection balance: sampling variation and response to subsequent directional selection**. *Genet Res* 1989, **54**(1):45-57.

309. Wright S: **Pleiotropy in the Evolution of Structural Reduction and of Dominance**. *American Naturalist* 1964, **98**(899):65-69.

310. Mackay TFC: **The genetic architecture of quantitative traits: lessons from Drosophila**. *Curr Opin Genet Dev* 2004, **14**(3):253-257.

311. Cohen BA, Mitra RD, Hughes JD, Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression**. *Nat Genet* 2000, **26**(2):183-186.

312. Pal C, Hurst LD: **Evidence for co-evolution of gene order and**

**recombination rate**. *Nature Genetics* 2003, **33**(3):392-395.

313. D'Ennequin ML, Toupance B, Robert T, Godelle B, Gouyon PH: **Plant domestication: a model for studying the selection of linkage**. *J Evolution Biol* 1999, **12**(6):1138-1147.
314. Nijhout HF: **Polymorphic mimicry in Papilio dardanus: mosaic dominance, big effects, and origins**. *Evol Dev* 2003, **5**(6):579-592.
315. Scarcelli N, Cheverud JM, Schaal BA, Kover PX: **Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus**. *Proc Natl Acad Sci U S A* 2007, **104**(43):16986-16991.
316. Jeffery WR: **Emerging model systems in evo-devo: cavefish and microevolution of development**. *Evol Dev* 2008, **10**(3):265-272.
317. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nat Rev Genet* 2011, **12**(2):87-98.
318. Draper BW, Morcos PA, Kimmel CB: **Inhibition of zebrafish fgf8 pre-mRNA splicing with Morpholino oligos: a quantifiable method for gene knockdown**. *Genesis* 2001, **30**(3):154–156.
319. Kopp A: **Metamodels and phylogenetic replication: a systematic approach to the evolution of developmental pathways**. *Evolution* 2009, **63**(11):2771-2789.

# Supplementary Material

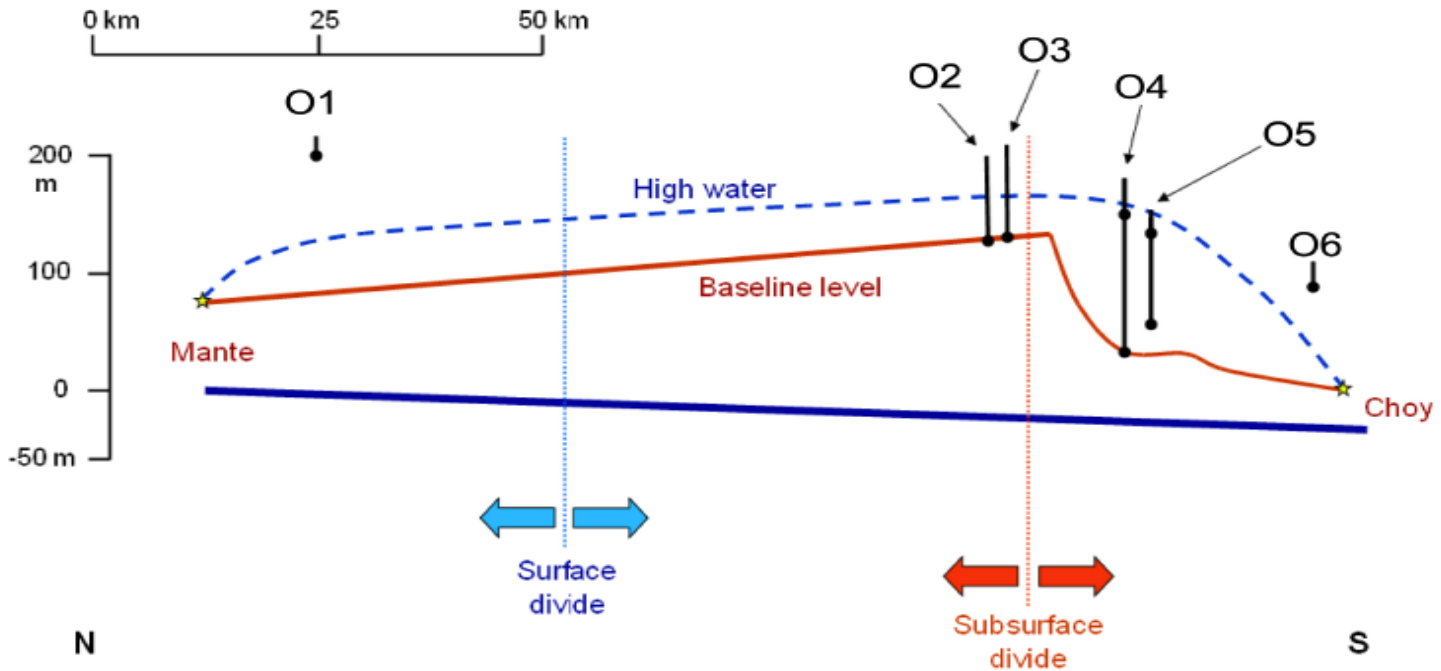
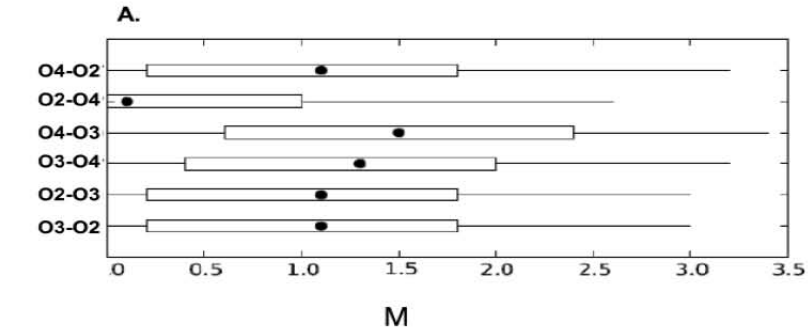
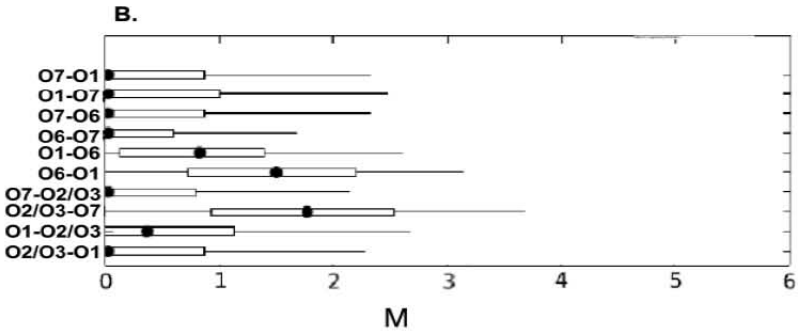
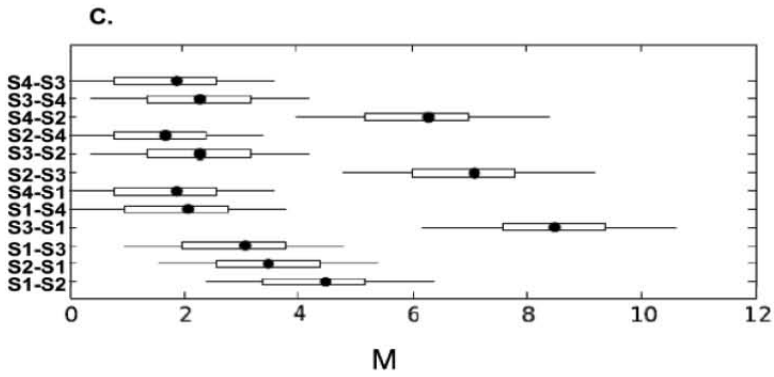
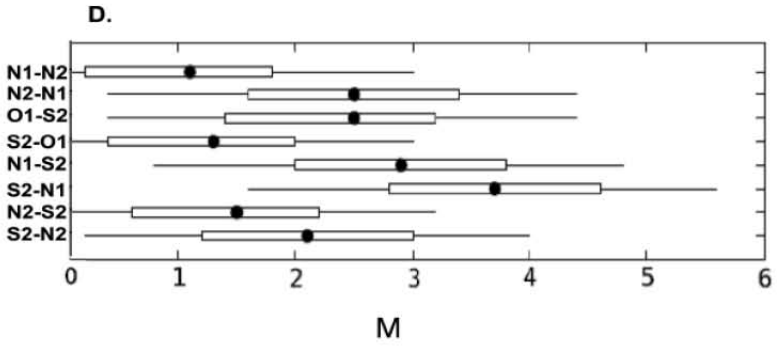



Figure S2.1. A detailed hydrological map of the El Abra region with the indication of surface and subsurface water divide. Points at, or near, base level (orange line) are indicated by solid circles; fish-inhabited pools by solid circles closer to the high water profile (blue dotted line) (adapted from Mitchell et al. 1977).


A.
O4-O2
O2-O4
O4-O3
O3-O4
O2-O3
O3-O2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
M


B.
O7-O1
O1-O7
O7-O6
O6-O7
O1-O6
O6-O1
O7-O2/O3
O2/O3-O7
O1-O2/O3
O2/O3-O1
0
1
2
3
4
5
6
M


C.
S4-S3
S3-S4
S4-S2
S2-S4
S3-S2
S2-S3
S4-S1
S1-S4
S3-S1
S1-S3
S2-S1
S1-S2
0
2
4
6
8
10
12
M


D.
N1-N2
N2-N1
O1-S2
S2-O1
N1-S2
S2-N1
N2-S2
S2-N2
0
1
2
3
4
5
6
M

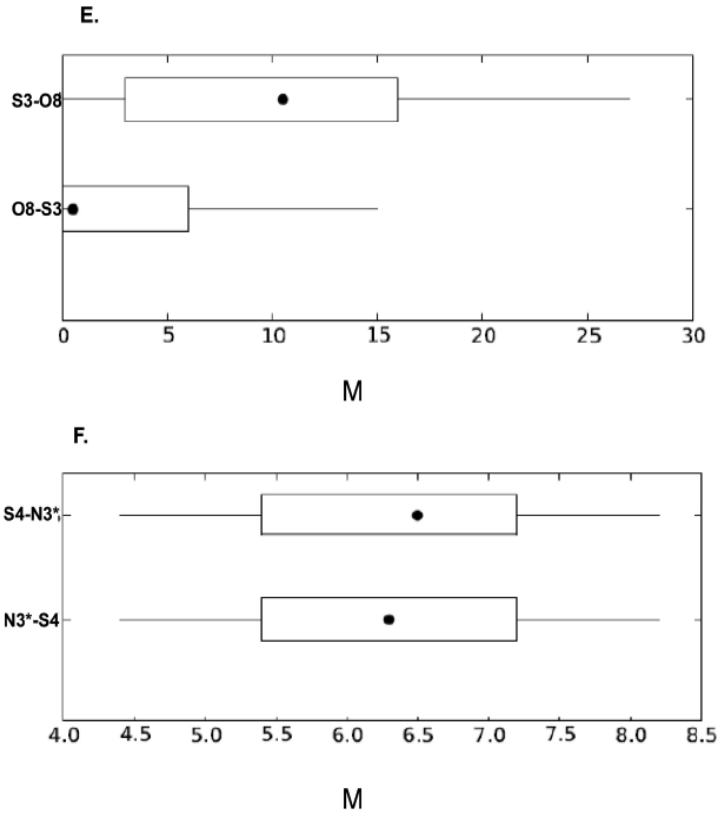



Figure S2.2. Estimates of gene flow based on Bayesian inferences of migration rates and population sizes using MIGRATE-N among *Astyanax mexicanus* population clusters within each geographical region. Mutation scaled immigration rate, M, between different population groups. M is the ratio of the immigration rate over the mutation rate. The central box of the plots represents the values from the lower to upper quartile (25 to 75 percentile). The middle dot represents the median posterior values over all loci. The horizontal line extends from the 2.5% percentile to the 97.5% percentile. Populations compared are designated to the left of the boxes (surface populations: S1-S4, new caves: N1-N3, old caves: O1-O8).

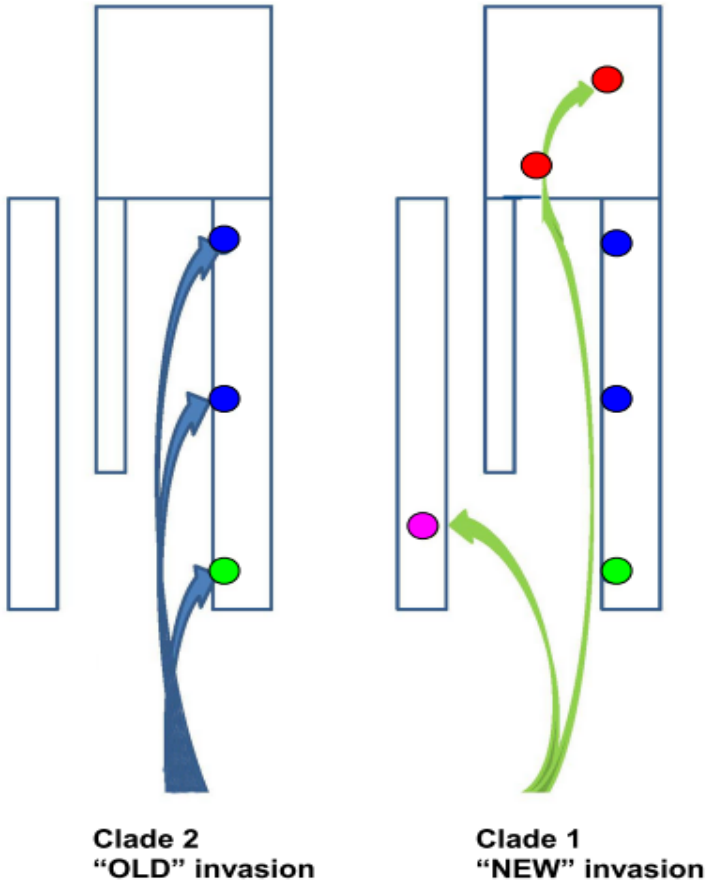



Figure S2.3. Summary of the proposed models. Proposed model with five independent origins of cave adapted *Astyanax* in NE Mexico as estimated by the data. The first wave of surface fish led to three independent subterranean invasion events establishing the "old" cave populations. The second wave gave rise to two independent invasions establishing "new" cave populations. The arrows signify that the ancestral stock moved into the area from the south but are not meant as specific routes.

Figure S3.1. SNP linkage map

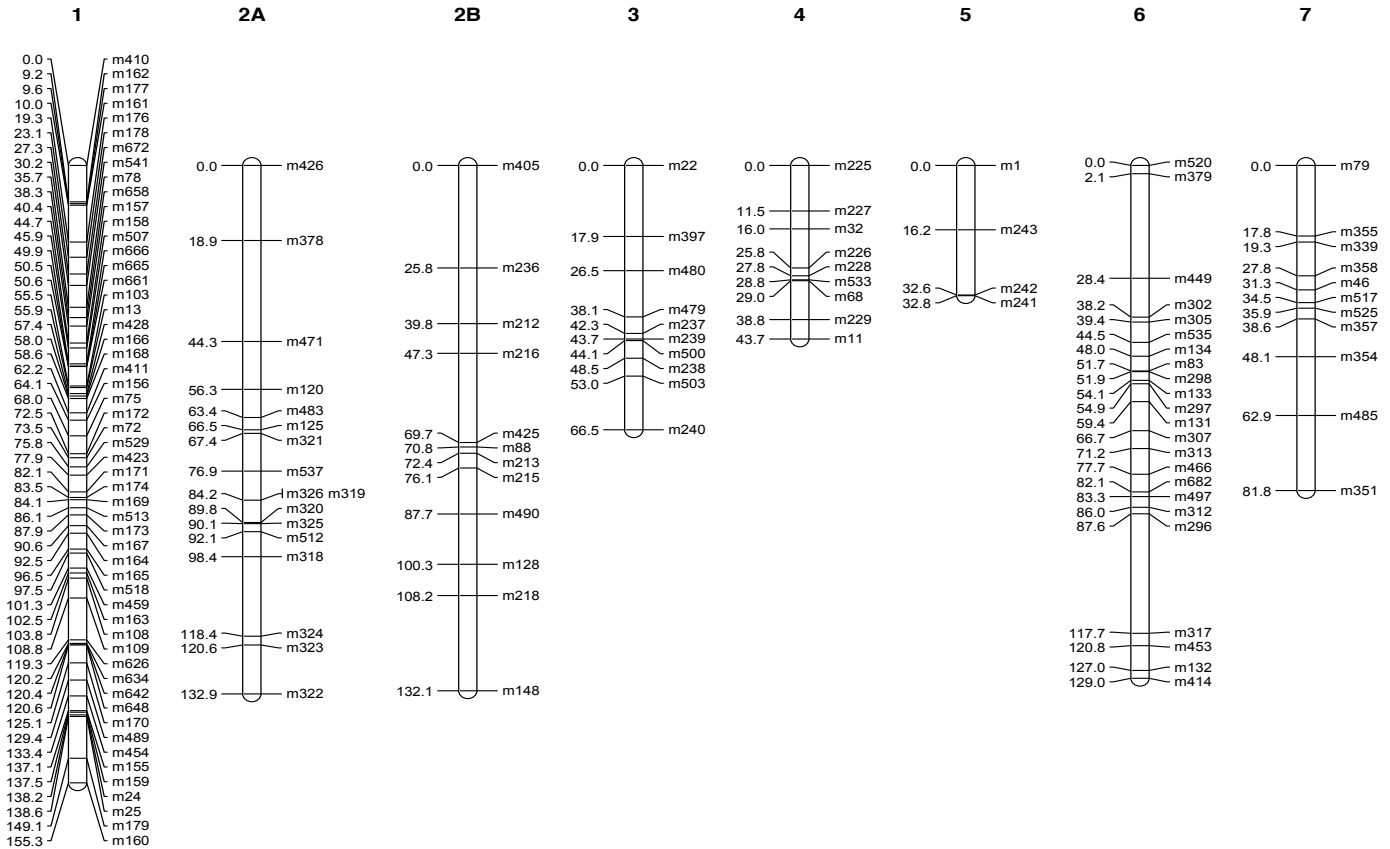


Figure S3.1. SNP linkage map

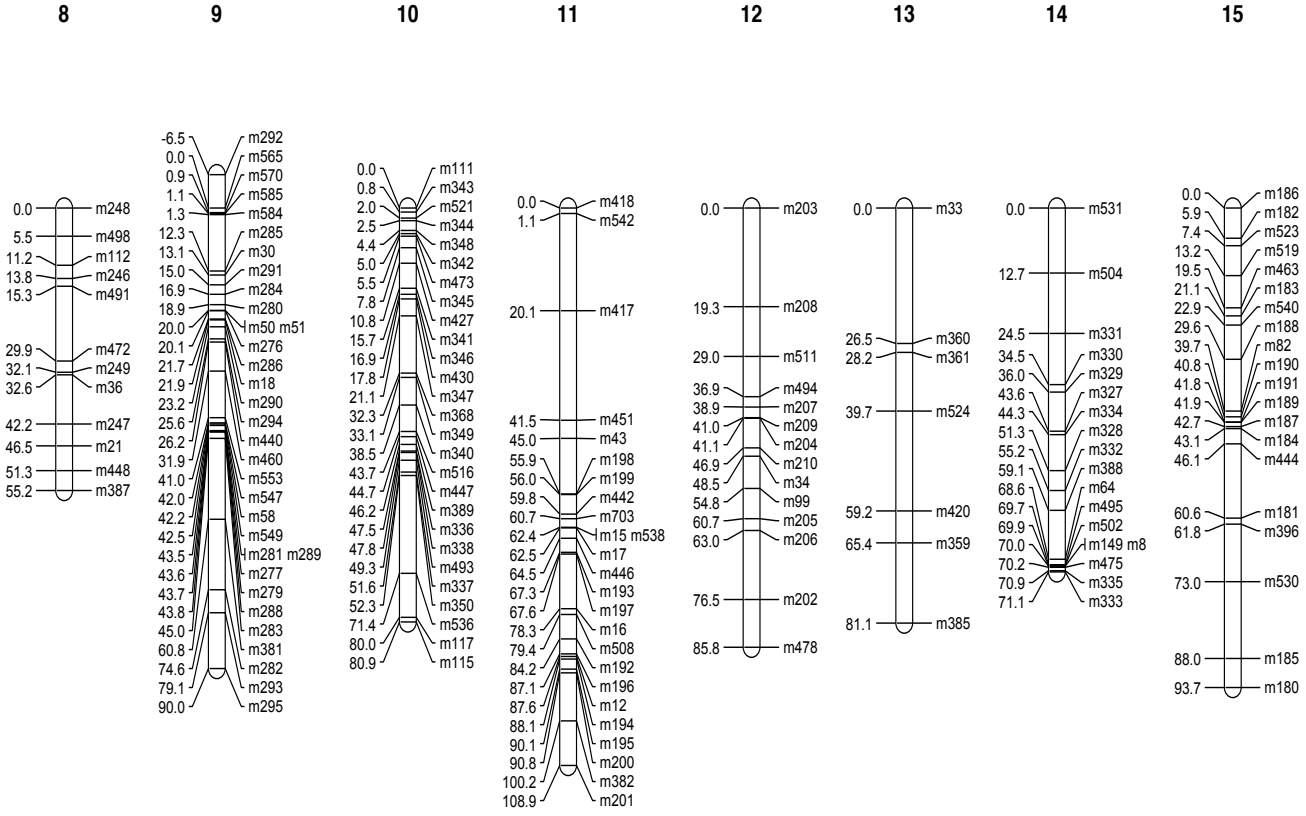


Figure S3.1. SNP linkage map

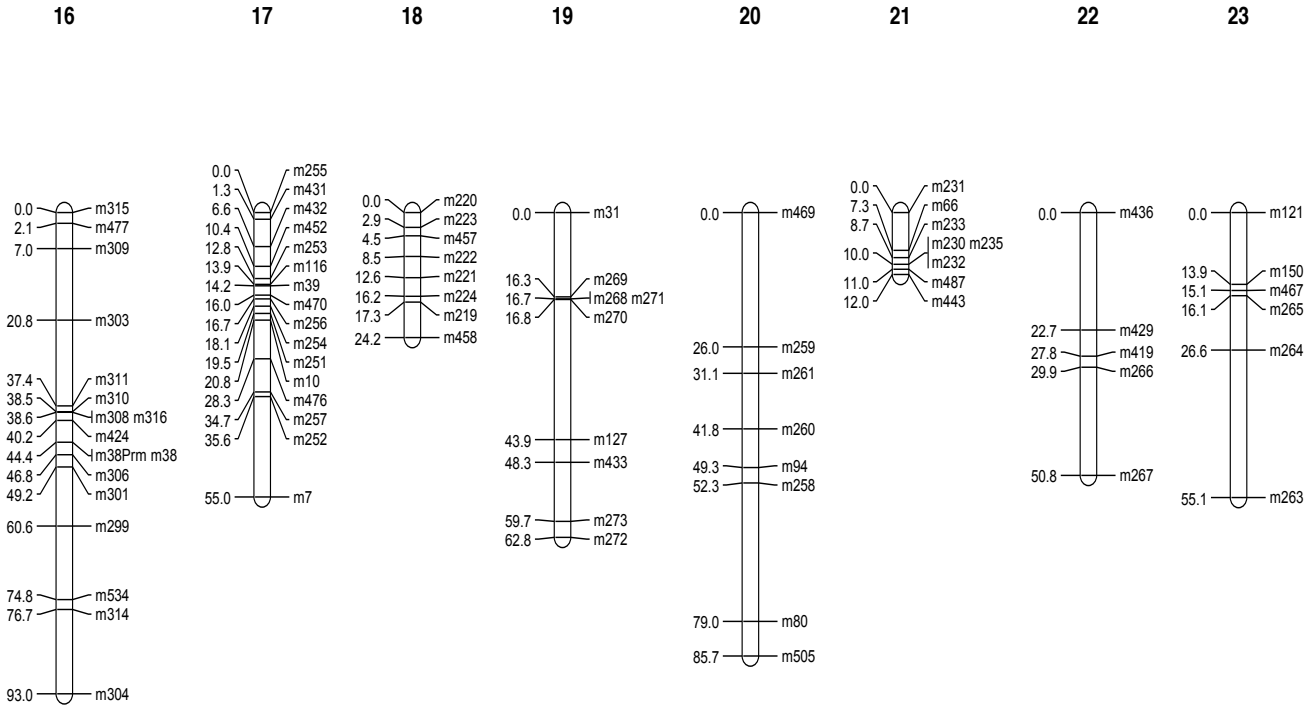



Figure S3.1. SNP only map of *Astyanax mexicanu*s with colored bars denoting positions of detected QTL for specific trait. Marker positions are given in centimorgans (cM).

Figure S3.2. Minor allele frequencies in each population

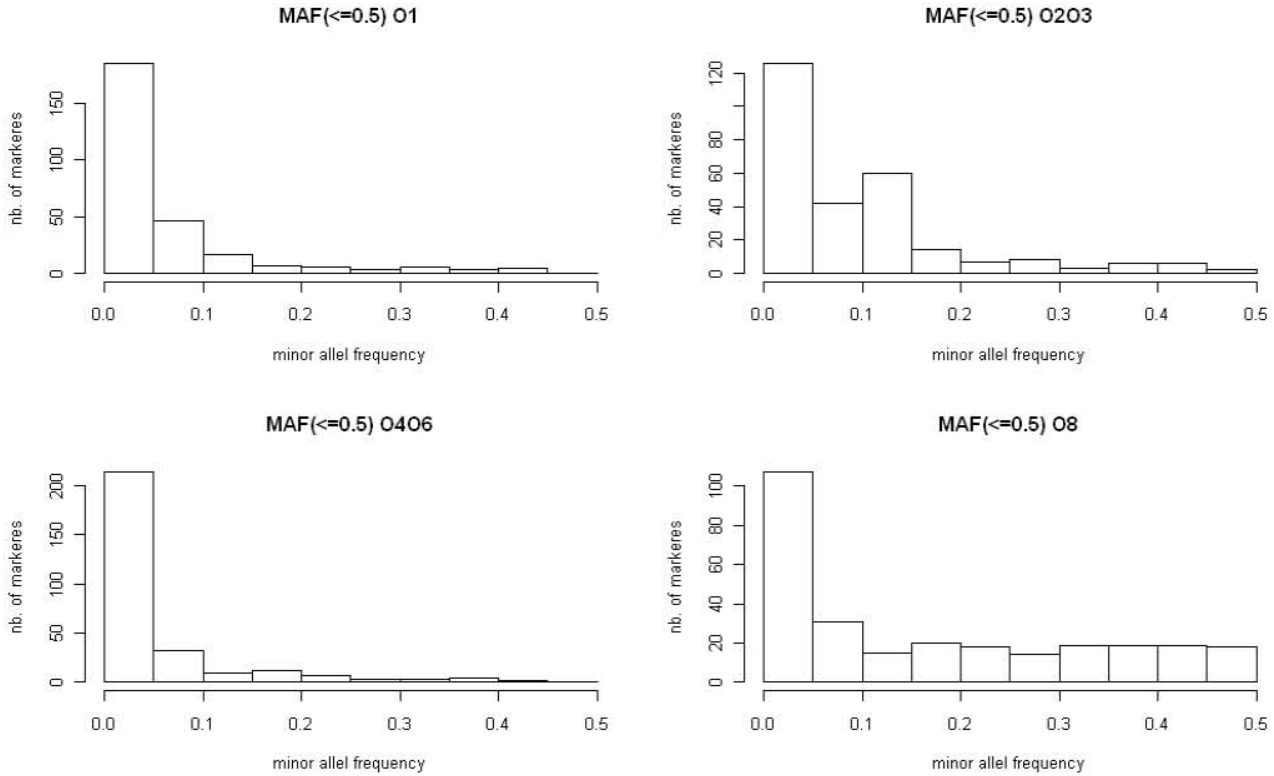
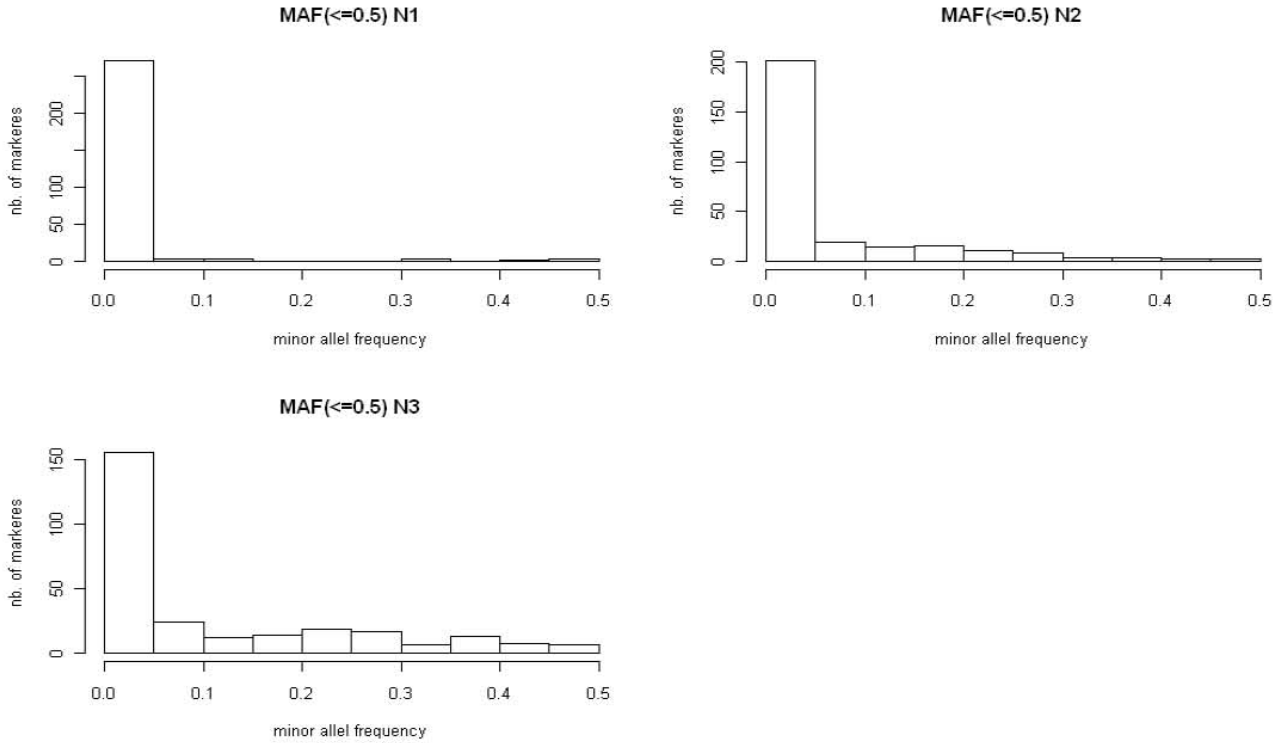


Figure S3.2. Minor allele frequencies in each population



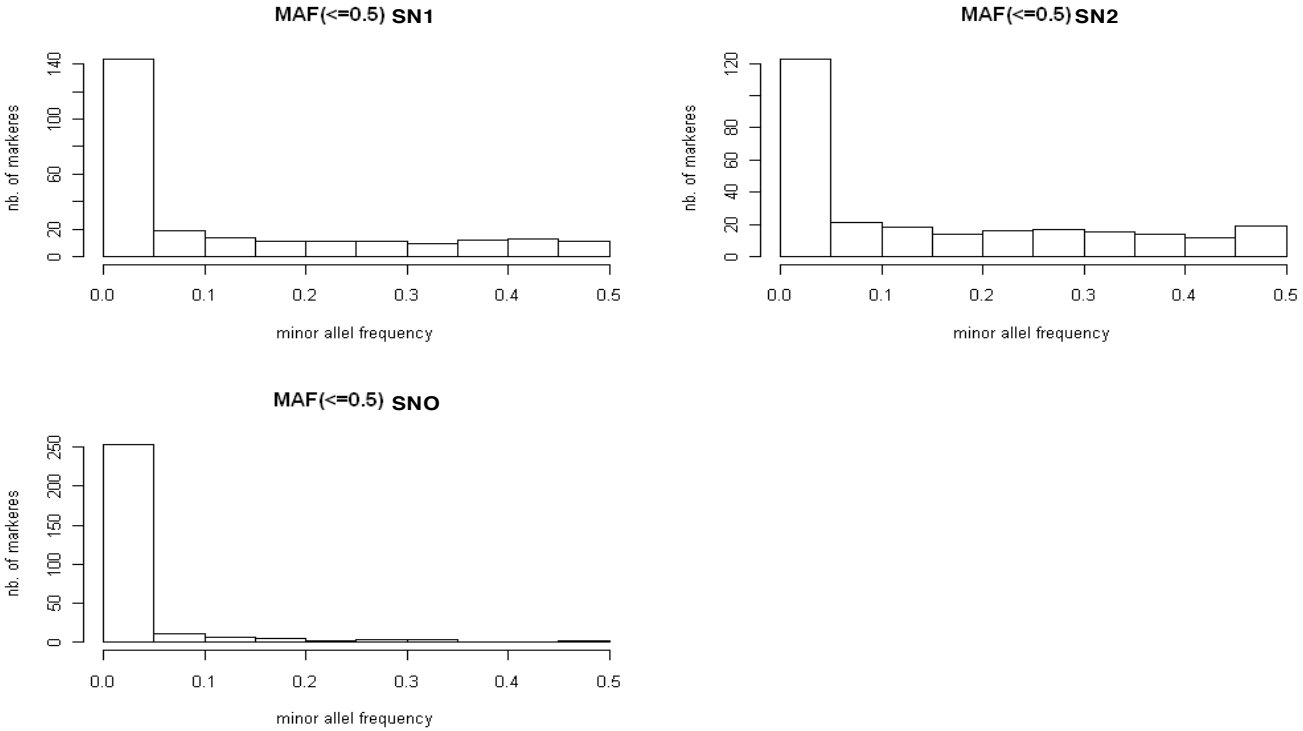



Figure S3.2. Minor allele frequencies in each population. Each figure represents each population with its acronyms. X-axis represents the frequency of the less common allele in the population (minor allele frequency with the frequency < 0.5). Y-axis represents number of markers per each MAF. Abbreviations of the population names are given above each figure.

| *TRAIT* | *LOD* | *L(cM)* | *P.E.V.* | *P.E.V.(ad)* |
|---|---|---|---|---|
| Eye_LG 1` | | | | |
| MV | 13.63 | 134.06 | 0.072 | 0.044 |
| CI | [ 7.0;20.2] | [120.5;147.6] | [0.04;0.11] | [0.01;0.07] |
| Eye _LG1 | | | | |
| MV | 6.519 | 88.38 | 0.035 | 0.03 |
| CI | [2.46;10.58] | [76.1;100.6] | [0.01;0.06] | [0.01;0.05] |
| MelA_LG2 | | | | |
| MV | 4.72 | 51.58 | 0.181 | 0.153 |
| CI | [0.54;8.90] | [24.0;79.2] | [0.01;0.35] | [0.02;0.29] |
| LensL_LG3 | | | | |
| MV | 8.815 | 103.18 | 0.134 | 0.057 |
| CI | [4.15;13.48] | [ 93.8;112.5] | [0.08;0.19] | [0.00;0.12] |
| RelDent_LG6 | | | | |
| MV | 5 | 76.02 | 0.069 | 0.0087 |
| CI | [1.41;8.60] | [55.2;85.9] | [0.02;0.12] | [0.00;0.03] |
| ResidLen_LG10 | | | | |
| MV | 5.114 | 16.85 | 0.065 | 0.054 |
| CI | [1.25;8.98] | [ 5.9;27.8] | [0.02;0.11] | [0.00;0.10] |
| Aasense_LG11 | | | | |
| MV | 5.954 | 8.442 | 0.115 | 0.055 |
| CI | [1.95;9.96] | [0.00;47.19] | [0.05;0.18] | [0.00;0.12] |
| Eye_LG12 | | | | |
| MV | 5.443 | 26.59 | 0.026 | 0.017 |
| CI | [1.65;9.24] | [ 0.0;74.7] | [0.01;0.04] | [0.00;0.03] |
| MEL_D_LG12 | | | | |
| MV | 4.645 | 46.18 | 0.242 | 0.162 |
| CI | [0.08;9.21] | [ 0.0;106.3] | [0.04;0.45] | [0.01;0.31] |
| Eye_LG14 | | | | |
| MV | 61.82 | 34.92 | 0.499 | 0.489 |
| CI | [48.7;74.9] | [33.3;36.5] | [0.44;0.55] | [0.43;0.55] |
| LensL_LG14 | | | | |
| MV | 11.05 | 65.34 | 0.169 | 0.158 |
| CI | [ 5.2;16.9] | [63.6;65.7] | [0.10;0.24] | [0.08;0.24] |
| Aasense_LG16 | | | | |
| MV | 4.616 | 12.07 | 0.08 | 0.009 |
| CI | [1.29;7.94] | [ 0.0;24.1] | [0.03;0.13] | [0.00;0.03] |
| LensE_LG20 | | | | |
| MV | 4.287 | 35.69 | 0.071 | 0.014 |
| CI | [1.73;6.84] | [23.3;46.8] | [0.02;0.12] | [0.00;0.04] |
| RelDent_LG25 | | | | |
| MV | 4.669 | 81.02 | 0.111 | 0.0084 |
| CI | [1.66;7.68] | [57.0;105.0] | [0.03;0.20] | [0.00;0.03] |
| Eye_LG27 | | | | |
| MV | 5.392 | 1.751 | 0.027 | 0.026 |
| CI | [0.80;9.99] | [0.00;6.37] | [0.00;0.05] | [0.00;0.05] |

| RelCond_LG28 | *LOD* | *L(cM)* | *P.E.V.* | *P.E.V.(ad)* |
|---|---|---|---|---|
| **MV** | 5.937 | 7.379 | 0.069 | 0.063 |
| **CI** | [1.44;10.44] | [0.00;29.73] | [0.02;0.12] | [0.01;0.11] |
| **WtLoss_LG28** | | | | |
| **MV** | 4.102 | 17.19 | 0.071 | 0.05 |
| **CI** | [1.17;7.03] | [ 0.0;35.1] | [0.00;0.14] | [0.00;0.12] |
| **Eye_LG29** | | | | |
| **MV** | 10.53 | 7.537 | 0.061 | 0.059 |
| **CI** | [ 4.2;16.9] | [0.00;27.16] | [0.03;0.09] | [0.03;0.09] |
| **Len_LG31** | | | | |
| **MV** | 4.454 | 3.32 | 0.08 | 0.011 |
| **CI** | [1.06;7.85] | [0.00;10.35] | [0.02;0.14] | [0.00;0.04] |
| **Length_LG34** | | | | |
| **MV** | 4.024 | 10.55 | 0.083 | 0.052 |
| **CI** | [0.66;7.39] | [ 0.0;23.5] | [0.02;0.15] | [0.00;0.11] |
| **Cond_LG33** | | | | |
| **MV** | 4.988 | 2.208 | 0.047 | 0.044 |
| **CI** | [1.25;8.73] | [0.00;5.92] | [0.01;0.08] | [0.01;0.08] |
| **WtLoss_LG33** | | | | |
| **MV** | 3.373 | 2.834 | 0.065 | 0.036 |
| **CI** | [0.00;6.97] | [1.85;3.82] | [0.00;0.13] | [0.00;0.09] |
| **Length_LG34** | | | | |
| **MV** | 4.024 | 10.55 | 0.083 | 0.052 |
| **CI** | [0.66;7.39] | [ 0.0;23.5] | [0.02;0.15] | [0.00;0.11] |
| **MelA_LG34** | | | | |
| **MV** | 4.2 | 16.63 | 0.088 | 0.043 |
| **CI(95%)** | [0.42;7.98] | [ 5.2;25.1] | [0.02;0.15] | [0.00;0.09] |
| **MelD_LG34** | | | | |
| **MV** | 3.708 | 14.65 | 0.039 | 0.014 |
| **CI** | [0.77;6.64] | [ 0.7;25.1] | [0.01;0.07] | [0.00;0.04] |
| **MelE_LG38** | | | | |
| **MV** | 4.314 | 19.94 | 0.062 | 0.036 |
| **CI** | [0.00;10.02] | [ 0.0;36.5] | [0.00;0.15] | [0.00;0.10] |

Table S3.1. Summary of identified QTL with their respective linkage group position and maximum LOD score. PEV and PEVad refer to the proportions of phenotypic trait variance in the mapping progeny ($F_2$) that are explained by a QTL. PEV refers to total trait variance; PEVad refers to the proportion of additive variance explained by the QTL. For each trait we are showing mean value (MV and confidence interval (CI, 95%) of each measure.

| QTL | M | NEW | | | OLD | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
| | m156 | 0.26 | 0.33 | 0.16 | 0.48 | 0.44 | 0.39 | 0.12 | + |
| | m172 | 0.09 | 0.05 | -1.00 | 0.16 | **0.01** | **0.00** | **0.01** | - |
| | m69 | 0.42 | 0.36 | 0.25 | 0.42 | 0.16 | 0.35 | 0.25 | + |
| | m75 | 0.43 | **0.00** | 0.14 | 0.31 | 0.20 | 0.33 | 0.42 | + |
| | m76 | 0.35 | 0.42 | 0.20 | 0.02 | 0.35 | 0.40 | 0.03 | - |
| | m70 | 0.27 | 0.21 | 0.23 | 0.13 | 0.27 | 0.22 | 0.12 | - |
| | m71 | 0.48 | 0.48 | 0.32 | 0.07 | 0.46 | 0.42 | 0.36 | + |
| | m72 | 0.22 | 0.02 | 0.23 | 0.07 | 0.04 | 0.24 | 0.35 | - |
| | m74 | 0.27 | 0.12 | 0.08 | **0.00** | 0.16 | 0.13 | 0.10 | + |
| | m529 | 0.07 | **0.03** | **0.01** | 0.12 | 0.15 | 0.12 | 0.05 | + |
| | m423 | -1.00 | -1.00 | 0.12 | 0.06 | -1.00 | -1.00 | -1.00 | - |
| | m171 | -1.00 | 0.05 | -1.00 | **0.00** | **0.03** | 0.05 | -1.00 | - |
| | m169 | -1.00 | -1.00 | -1.00 | **0.03** | -1.00 | -1.00 | -1.00 | - |
| | m174 | 0.20 | 0.13 | **0.04** | 0.32 | 0.41 | 0.29 | 0.25 | + |
| | m513 | 0.19 | 0.15 | 0.11 | 0.23 | 0.22 | 0.22 | 0.10 | + |
| | m173 | -1.00 | -1.00 | -1.00 | 0.06 | -1.00 | -1.00 | -1.00 | - |
| | m167 | 0.30 | 0.17 | 0.20 | 0.39 | 0.44 | 0.46 | 0.46 | + |
| RelEye | m164 | 0.08 | **0.04** | -1.00 | 0.05 | **0.02** | **0.02** | **0.04** | - |
| | m165 | -1.00 | -1.00 | -1.00 | 0.06 | **0.03** | -1.00 | -1.00 | - |
| | m518 | -1.00 | -1.00 | **0.04** | 0.04 | **0.02** | 0.06 | 0.02 | - |
| | m459 | -1.00 | -1.00 | 0.17 | **0.02** | **0.01** | 0.05 | 0.03 | - |
| | m163 | 0.08 | 0.08 | 0.13 | 0.10 | 0.06 | **0.03** | 0.33 | - |
| | m108 | 0.13 | **0.03** | 0.13 | 0.30 | 0.20 | 0.23 | 0.22 | + |
| | m109 | 0.17 | **0.04** | 0.18 | 0.34 | 0.28 | 0.26 | 0.24 | + |
| | m177 | 0.41 | 0.06 | 0.42 | 0.25 | 0.40 | 0.42 | 0.38 | + |
| | m176 | 0.07 | **0.02** | 0.05 | 0.06 | 0.25 | 0.04 | 0.25 | - |
| | m178 | 0.28 | 0.32 | 0.34 | 0.43 | 0.45 | 0.41 | 0.42 | + |
| | m541 | 0.08 | 0.06 | 0.12 | 0.11 | 0.15 | 0.15 | 0.08 | + |
| | m507 | -1.00 | -1.00 | **0.04** | **0.03** | 0.08 | **0.04** | 0.06 | - |
| | m78 | -1.00 | -1.00 | -1.00 | 0.07 | **0.02** | **0.00** | **0.00** | - |
| | m669 | **0.02** | **0.01** | 0.08 | **0.00** | **0.03** | **0.01** | **0.00** | - |
| | m670 | 0.21 | 0.13 | 0.27 | 0.40 | 0.45 | 0.36 | 0.14 | + |
| | m663 | 0.47 | 0.40 | **0.00** | 0.45 | 0.47 | 0.45 | 0.29 | + |
| | m660 | 0.41 | 0.18 | 0.38 | 0.34 | 0.38 | 0.34 | **0.02** | + |
| | m661 | 0.19 | 0.11 | 0.26 | **0.04** | **0.02** | 0.11 | **0.01** | - |
| | m662 | 0.21 | 0.39 | 0.35 | **0.01** | 0.30 | 0.43 | 0.12 | + |
| | m665 | 0.23 | 0.26 | 0.26 | 0.44 | 0.41 | 0.37 | 0.19 | + |
| | m666 | 0.22 | 0.27 | 0.25 | 0.09 | **0.04** | 0.06 | **0.01** | - |
| RelEye | m667 | 0.44 | 0.04 | 0.10 | 0.45 | 0.24 | 0.36 | 0.28 | + |
| | m168 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | -1.00 | -1.00 | - |
| | m166 | -1.00 | -1.00 | -1.00 | **0.01** | -1.00 | -1.00 | -1.00 | - |
| | m428 | 0.21 | 0.27 | 0.15 | **0.02** | **0.03** | 0.24 | 0.34 | - |
| | m47 | 0.29 | 0.24 | 0.40 | 0.43 | 0.47 | 0.42 | 0.23 | + |
| | m48 | 0.35 | 0.32 | 0.15 | 0.39 | 0.39 | 0.35 | 0.27 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m49 | 0.39 | 0.21 | **0.02** | 0.19 | 0.05 | 0.35 | 0.41 | + |
| | m50 | -1.00 | -1.00 | -1.00 | 0.07 | **0.02** | **0.03** | -1.00 | - |
| | m51 | -1.00 | -1.00 | -1.00 | 0.07 | **0.02** | **0.03** | -1.00 | - |
| | m53 | 0.24 | 0.17 | 0.42 | 0.46 | 0.38 | 0.46 | 0.20 | + |
| | m52 | 0.48 | 0.43 | 0.29 | 0.49 | 0.48 | 0.47 | 0.32 | + |
| | m291 | 0.05 | **0.04** | 0.26 | 0.05 | 0.07 | **0.04** | **0.01** | - |
| | m285 | 0.45 | 0.38 | 0.32 | 0.46 | 0.06 | 0.49 | 0.34 | + |
| | m292 | 0.12 | 0.09 | 0.09 | **0.04** | 0.05 | **0.03** | **0.01** | + |
| | m565 | **0.01** | 0.32 | 0.25 | **0.03** | **0.04** | 0.06 | **0.03** | - |
| | m566 | 0.26 | 0.19 | 0.15 | 0.42 | 0.47 | 0.35 | 0.28 | + |
| | m564 | -1.00 | 0.05 | -1.00 | -1.00 | -1.00 | **0.04** | **0.02** | - |
| | m567 | 0.44 | **0.03** | 0.19 | 0.42 | 0.28 | 0.36 | **0.02** | + |
| | m568 | 0.42 | **0.02** | 0.19 | 0.41 | 0.15 | 0.34 | **0.01** | + |
| | m569 | 0.22 | 0.27 | 0.25 | 0.33 | 0.22 | 0.26 | 0.35 | - |
| | m587 | 0.22 | 0.26 | 0.25 | 0.33 | 0.22 | 0.26 | 0.34 | - |
| | m588 | 0.46 | 0.37 | 0.30 | 0.21 | 0.30 | 0.34 | 0.23 | + |
| | m589 | 0.42 | 0.36 | 0.23 | 0.47 | 0.26 | 0.47 | 0.41 | + |
| | m570 | 0.45 | **0.03** | 0.22 | 0.40 | 0.29 | 0.06 | 0.28 | + |
| | m571 | 0.42 | 0.36 | 0.23 | 0.47 | 0.26 | 0.47 | 0.41 | + |
| | m572 | 0.39 | 0.33 | 0.20 | 0.40 | 0.33 | 0.41 | 0.39 | + |
| | m573 | 0.31 | 0.18 | **0.02** | 0.41 | 0.30 | 0.33 | 0.33 | + |
| | m574 | 0.30 | 0.36 | 0.45 | 0.28 | 0.29 | 0.27 | **0.02** | + |
| | m577 | 0.35 | 0.27 | 0.35 | 0.47 | 0.26 | **0.01** | 0.28 | + |
| | m579 | 0.32 | 0.37 | 0.38 | 0.45 | 0.32 | 0.07 | **0.00** | + |
| | m584 | 0.18 | 0.24 | 0.40 | 0.48 | 0.46 | 0.22 | 0.12 | + |
| | m585 | 0.18 | 0.24 | 0.46 | 0.48 | 0.46 | 0.22 | 0.12 | + |
| | m591 | **0.03** | 0.17 | 0.07 | 0.17 | 0.17 | 0.14 | **0.02** | + |
| MelA | m592 | 0.38 | 0.16 | 0.36 | 0.43 | 0.09 | 0.33 | 0.10 | + |
| | m283 | 0.25 | 0.17 | 0.20 | 0.37 | 0.43 | 0.34 | 0.30 | + |
| | m279 | 0.05 | **0.04** | -1.00 | 0.13 | **0.02** | 0.06 | **0.01** | - |
| | m277 | 0.26 | 0.38 | 0.25 | 0.43 | 0.44 | 0.15 | 0.26 | - |
| | m281 | **0.00** | **0.00** | -1.00 | **0.02** | **0.02** | **0.04** | **0.02** | - |
| | m289 | 0.40 | 0.08 | 0.14 | 0.44 | 0.29 | 0.31 | 0.47 | + |
| | m288 | 0.15 | 0.12 | 0.19 | 0.34 | 0.23 | 0.24 | 0.21 | + |
| | m545 | 0.42 | 0.37 | 0.25 | 0.44 | 0.43 | 0.45 | 0.18 | + |
| | m543 | 0.47 | 0.40 | 0.30 | 0.45 | 0.47 | 0.47 | 0.16 | + |
| | m544 | 0.29 | 0.16 | 0.16 | 0.41 | 0.32 | 0.38 | 0.29 | + |
| | m546 | 0.40 | 0.12 | 0.15 | 0.25 | 0.20 | 0.18 | 0.19 | + |
| | m547 | 0.46 | 0.43 | 0.34 | 0.13 | 0.13 | 0.09 | 0.08 | + |
| | m549 | 0.37 | 0.18 | 0.22 | 0.36 | 0.26 | 0.28 | 0.27 | + |
| | m553 | 0.21 | 0.09 | **0.02** | 0.11 | 0.13 | 0.25 | 0.08 | + |
| | m550 | 0.22 | 0.28 | 0.42 | 0.41 | 0.33 | 0.29 | 0.43 | + |
| | m552 | 0.42 | 0.35 | 0.37 | 0.48 | 0.37 | 0.35 | 0.27 | + |
| | m557 | 0.48 | 0.06 | 0.37 | 0.14 | 0.31 | 0.43 | **0.01** | + |
| | m563 | 0.22 | 0.26 | 0.25 | 0.33 | 0.22 | 0.26 | 0.34 | - |
| | m554 | 0.46 | 0.36 | 0.34 | 0.43 | 0.22 | 0.34 | 0.11 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m556 | 0.25 | 0.29 | **0.02** | 0.21 | 0.17 | 0.23 | 0.11 | + |
| RelEye | m201 | 0.22 | 0.13 | 0.34 | 0.48 | 0.23 | 0.20 | 0.05 | + |
| | m197 | -1.00 | -1.00 | -1.00 | **0.00** | -1.00 | -1.00 | -1.00 | - |
| | m193 | 0.22 | 0.27 | 0.25 | **0.04** | 0.05 | 0.23 | 0.35 | - |
| | m446 | 0.23 | 0.22 | 0.12 | 0.23 | 0.22 | 0.23 | 0.14 | + |
| | m15 | -1.00 | -1.00 | -1.00 | **0.00** | 0.08 | 0.11 | 0.12 | - |
| | m538 | 0.32 | 0.45 | 0.34 | 0.49 | 0.37 | 0.44 | 0.40 | + |
| | m17 | -1.00 | -1.00 | -1.00 | **0.00** | 0.05 | **0.00** | **0.01** | - |
| | m703 | **0.01** | **0.01** | **0.00** | 0.09 | **0.04** | **0.02** | **0.01** | - |
| Lens L | m702 | 0.26 | **0.04** | 0.07 | 0.39 | **0.01** | 0.31 | 0.42 | + |
| | m442 | 0.05 | **0.04** | **0.03** | 0.49 | 0.49 | 0.36 | 0.36 | + |
| | m198 | -1.00 | -1.00 | -1.00 | **0.01** | -1.00 | -1.00 | -1.00 | - |
| | m199 | 0.18 | 0.03 | 0.32 | 0.10 | 0.08 | 0.14 | 0.09 | + |
| | m43 | 0.37 | 0.31 | 0.44 | 0.39 | 0.45 | 0.39 | 0.33 | + |
| | m418 | 0.38 | 0.16 | **0.04** | 0.46 | 0.15 | 0.45 | **0.03** | + |
| WtLoss | m542 | 0.31 | 0.29 | 0.20 | 0.48 | 0.38 | 0.49 | 0.08 | + |
| | m213 | -1.00 | -1.00 | -1.00 | **0.03** | -1.00 | -1.00 | -1.00 | - |
| | m88 | **0.04** | **0.02** | 0.06 | 0.40 | 0.23 | 0.31 | 0.45 | + |
| | m425 | **0.01** | **0.02** | 0.23 | 0.05 | 0.08 | **0.02** | 0.12 | - |
| | m215 | 0.21 | 0.06 | **0.04** | **0.01** | 0.28 | 0.18 | 0.25 | + |
| | m218 | -1.00 | -1.00 | -1.00 | 0.02 | 0.04 | 0.14 | -1.00 | - |
| AASens | m211 | 0.05 | 0.05 | -1.00 | **0.01** | **0.03** | 0.05 | **0.01** | - |
| | m386 | 0.47 | 0.38 | 0.26 | 0.41 | 0.50 | 0.42 | 0.24 | + |
| | m270 | 0.27 | 0.45 | 0.06 | 0.49 | 0.37 | 0.46 | 0.45 | + |
| | m268 | -1.00 | -1.00 | -1.00 | 0.12 | -1.00 | -1.00 | -1.00 | - |
| | m269 | 0.41 | 0.06 | 0.42 | 0.15 | 0.11 | 0.06 | 0.28 | + |
| | m271 | 0.21 | 0.25 | **0.04** | 0.50 | 0.31 | 0.31 | 0.44 | + |
| | m127 | 0.28 | 0.19 | 0.05 | 0.27 | 0.28 | 0.26 | 0.29 | + |
| | m433 | 0.39 | 0.43 | 0.11 | 0.23 | 0.38 | 0.18 | 0.32 | + |
| | m273 | 0.26 | 0.17 | 0.48 | 0.05 | 0.40 | 0.43 | 0.34 | + |
| | m272 | 0.10 | 0.13 | **0.03** | 0.15 | 0.10 | 0.13 | 0.06 | + |
| | m180 | 0.44 | 0.32 | 0.05 | 0.41 | 0.47 | 0.48 | 0.42 | + |
| | m185 | 0.27 | 0.24 | **0.02** | 0.30 | 0.21 | 0.22 | **0.02** | + |
| | m181 | 0.33 | 0.41 | 0.38 | **0.01** | 0.27 | 0.33 | 0.18 | - |
| RelDent | m444 | -1.00 | -1.00 | 0.08 | **0.02** | 0.02 | 0.05 | **0.00** | - |
| | m184 | 0.22 | 0.14 | 0.09 | 0.49 | 0.40 | 0.35 | 0.24 | + |
| | m187 | 0.05 | 0.23 | -1.00 | **0.01** | **0.02** | 0.06 | **0.00** | - |
| | m189 | 0.05 | 0.03 | 0.25 | 0.17 | 0.13 | 0.19 | 0.20 | + |
| RelDent | m191 | 0.23 | **0.02** | 0.32 | 0.06 | 0.08 | 0.12 | 0.15 | + |
| | m82 | 0.05 | 0.17 | **0.04** | 0.08 | 0.10 | 0.15 | **0.04** | + |
| | m190 | 0.36 | 0.24 | 0.15 | 0.44 | 0.47 | 0.47 | **0.00** | + |
| RelEye | m188 | 0.38 | **0.02** | 0.30 | 0.16 | 0.07 | 0.29 | 0.35 | + |
| | m183 | 0.22 | 0.35 | 0.08 | 0.46 | 0.08 | 0.45 | 0.28 | + |
| | m540 | 0.30 | 0.19 | 0.06 | 0.48 | 0.39 | 0.43 | **0.03** | + |
| | m463 | 0.32 | 0.21 | 0.04 | **0.03** | 0.43 | **0.04** | 0.26 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m519 | -1.00 | -1.00 | **0.04** | **0.00** | -1.00 | -1.00 | -1.00 | - |
| | m523 | 0.11 | 0.44 | 0.14 | 0.19 | 0.42 | 0.44 | 0.31 | + |
| | m182 | 0.30 | 0.47 | 0.13 | 0.49 | 0.48 | 0.45 | 0.05 | + |
| | m115 | 0.30 | 0.35 | 0.35 | 0.25 | 0.42 | **0.00** | 0.41 | + |
| | m117 | 0.41 | **0.01** | 0.14 | 0.15 | 0.09 | 0.09 | 0.24 | + |
| | m536 | -1.00 | -1.00 | **0.04** | **0.03** | 0.14 | -1.00 | -1.00 | - |
| | m350 | 0.35 | 0.17 | 0.23 | 0.47 | 0.42 | 0.49 | 0.44 | + |
| | m337 | 0.41 | 0.21 | 0.18 | 0.35 | 0.44 | 0.41 | **0.04** | + |
| | m338 | 0.08 | **0.04** | -1.00 | **0.02** | **0.03** | 0.06 | **0.02** | - |
| | m336 | 0.05 | 0.05 | -1.00 | 0.07 | **0.02** | 0.07 | **0.01** | - |
| | m493 | -1.00 | 0.22 | 0.08 | **0.02** | **0.02** | **0.04** | **0.02** | - |
| | m389 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | -1.00 | -1.00 | - |
| | m447 | -1.00 | -1.00 | 0.16 | **0.01** | **0.02** | **0.04** | -1.00 | - |
| | m516 | -1.00 | -1.00 | 0.15 | 0.08 | -1.00 | -1.00 | -1.00 | - |
| | m340 | 0.19 | 0.37 | 0.28 | 0.35 | 0.39 | 0.35 | 0.31 | + |
| | m347 | -1.00 | -1.00 | 0.13 | **0.00** | 0.18 | -1.00 | -1.00 | - |
| | m430 | 0.44 | 0.41 | **0.02** | 0.48 | **0.01** | 0.45 | 0.08 | + |
| | m346 | **0.00** | 0.05 | -1.00 | **0.00** | **0.00** | **0.04** | **0.00** | - |
| | m341 | 0.42 | 0.09 | 0.09 | **0.04** | 0.07 | 0.18 | 0.21 | - |
| | m427 | 0.24 | 0.34 | 0.44 | 0.36 | 0.36 | 0.34 | 0.35 | + |
| | m473 | 0.19 | 0.07 | 0.10 | 0.37 | 0.37 | 0.36 | 0.42 | + |
| | m521 | 0.31 | 0.38 | 0.07 | 0.46 | 0.43 | 0.42 | 0.18 | + |
| | m344 | 0.09 | 0.07 | 0.10 | 0.13 | **0.01** | 0.28 | 0.14 | + |
| | m348 | 0.12 | 0.40 | **0.02** | 0.16 | 0.19 | 0.22 | 0.09 | + |
| | m342 | -1.00 | -1.00 | -1.00 | **0.03** | **0.03** | -1.00 | -1.00 | - |
| | m345 | **0.01** | **0.00** | **0.00** | **0.00** | **0.02** | **0.01** | **0.00** | - |
| COND | m331 | 0.25 | 0.32 | 0.07 | 0.05 | 0.26 | 0.31 | 0.42 | + |
| MelA | m329 | -1.00 | -1.00 | -1.00 | **0.03** | 0.09 | -1.00 | -1.00 | - |
| | m327 | 0.37 | 0.43 | 0.28 | 0.30 | 0.20 | 0.30 | 0.19 | + |
| | m334 | 0.05 | 0.13 | 0.09 | 0.20 | 0.05 | 0.09 | 0.21 | - |
| | m328 | 0.15 | **0.04** | 0.30 | 0.31 | 0.21 | 0.32 | 0.28 | + |
| | m335 | 0.22 | **0.01** | 0.25 | 0.09 | 0.22 | 0.26 | 0.37 | - |
| | m333 | 0.21 | 0.29 | **0.04** | 0.35 | 0.23 | 0.28 | 0.31 | - |
| | m502 | -1.00 | -1.00 | 0.16 | **0.04** | -1.00 | -1.00 | -1.00 | - |
| | m495 | 0.21 | 0.25 | 0.10 | 0.40 | 0.40 | 0.40 | 0.30 | + |
| | m8 | -1.00 | -1.00 | -1.00 | 0.05 | 0.16 | -1.00 | -1.00 | - |
| | m64 | -1.00 | -1.00 | -1.00 | **0.00** | **0.02** | 0.08 | 0.09 | - |
| | m21 | -1.00 | -1.00 | -1.00 | **0.02** | **0.02** | -1.00 | -1.00 | - |
| | m247 | 0.33 | 0.49 | **0.03** | 0.43 | 0.17 | 0.50 | 0.34 | + |
| | m30 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | -1.00 | -1.00 | - |
| | m284 | 0.09 | 0.05 | -1.00 | **0.02** | **0.02** | 0.06 | **0.02** | - |
| | m280 | 0.09 | 0.05 | -1.00 | **0.02** | **0.01** | 0.06 | **0.01** | - |
| | m276 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | -1.00 | -1.00 | - |
| | m605 | 0.14 | 0.41 | **0.03** | 0.16 | 0.22 | 0.48 | 0.19 | + |
| | m606 | 0.33 | 0.41 | 0.07 | 0.30 | 0.32 | 0.30 | 0.41 | + |
| ResidLen | m598 | 0.37 | 0.27 | 0.39 | **0.03** | 0.24 | 0.14 | 0.19 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m379 | 0.05 | **0.02** | 0.09 | 0.10 | 0.07 | 0.27 | 0.08 | + |
| | m485 | -1.00 | -1.00 | 0.15 | **0.03** | -1.00 | -1.00 | -1.00 | - |
| | m691 | -1 | -1 | -1 | -1 | **0.03** | -1 | -1 | - |
| MEI E | m692 | -1 | -1 | -1 | -1 | 0.17 | -1 | -1 | - |
| | m357 | -1.00 | -1.00 | -1.00 | 0.08 | **0.03** | -1.00 | -1.00 | - |
| | m525 | **0.03** | **0.01** | 0.33 | 0.17 | 0.08 | 0.08 | **0.03** | - |
| | m517 | -1.00 | -1.00 | **0.04** | **0.03** | **0.02** | **0.04** | -1.00 | - |
| | m46 | 0.43 | **0.03** | 0.16 | **0.03** | 0.39 | 0.37 | 0.22 | + |
| | m358 | 0.10 | **0.05** | -1.00 | 0.05 | **0.02** | **0.03** | **0.01** | - |
| | m339 | 0.36 | 0.37 | 0.44 | 0.39 | 0.49 | 0.38 | 0.06 | + |
| | m355 | 0.24 | 0.20 | 0.35 | 0.25 | 0.44 | 0.22 | 0.09 | + |
| | m303 | 0.38 | **0.02** | 0.10 | 0.43 | **0.02** | 0.47 | 0.36 | + |
| | m309 | -1.00 | -1.00 | -1.00 | **0.02** | 0.09 | -1.00 | -1.00 | - |
| | m315 | **0.03** | **0.04** | **0.04** | 0.12 | 0.12 | 0.07 | 0.16 | + |
| Eye | m477 | 0.27 | 0.27 | 0.48 | **0.02** | 0.29 | 0.37 | 0.29 | + |
| | m414 | 0.07 | 0.05 | -1.00 | **0.00** | **0.02** | **0.04** | **0.02** | - |
| | m132 | 0.23 | 0.14 | 0.38 | 0.43 | 0.29 | 0.33 | 0.13 | + |
| | m453 | 0.29 | 0.21 | 0.43 | 0.45 | 0.49 | 0.12 | 0.09 | + |
| | m317 | 0.06 | 0.16 | 0.11 | 0.10 | 0.36 | 0.14 | 0.25 | - |
| | m25 | -1.00 | -1.00 | -1.00 | 0.08 | -1.00 | -1.00 | -1.00 | - |
| | m24 | 0.21 | **0.04** | 0.11 | 0.34 | 0.41 | 0.32 | 0.26 | + |
| | m155 | **0.01** | 0.11 | 0.24 | 0.22 | 0.23 | 0.24 | 0.36 | - |
| | m159 | 0.22 | 0.14 | 0.25 | 0.10 | 0.06 | **0.02** | 0.21 | - |
| | m454 | -1.00 | -1.00 | **0.04** | **0.04** | 0.01 | 0.20 | -1.00 | - |
| | m489 | 0.37 | 0.32 | 0.19 | 0.08 | 0.39 | 0.42 | 0.26 | + |
| | m170 | 0.46 | **0.03** | 0.18 | 0.05 | 0.17 | 0.30 | 0.44 | + |
| | m626 | 0.42 | 0.32 | 0.26 | 0.48 | 0.35 | 0.35 | 0.40 | - |
| | m627 | 0.32 | 0.37 | 0.37 | 0.46 | 0.32 | 0.35 | 0.15 | + |
| | m628 | 0.34 | 0.05 | 0.24 | 0.40 | 0.35 | 0.49 | 0.24 | + |
| AASens | m629 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | - |
| | m630 | 0.27 | 0.19 | 0.23 | 0.13 | 0.27 | 0.22 | 0.11 | + |
| | m637 | **0.02** | **0.00** | 0.10 | 0.13 | **0.03** | 0.12 | **0.02** | - |
| | m639 | 0.36 | 0.16 | 0.24 | 0.34 | 0.41 | 0.35 | 0.22 | + |
| | m642 | 0.46 | 0.11 | 0.18 | 0.37 | 0.33 | 0.32 | 0.15 | + |
| | m632 | 0.42 | 0.18 | **0.00** | **0.04** | **0.03** | 0.33 | 0.07 | - |
| | m634 | 0.46 | 0.40 | **0.03** | 0.38 | 0.33 | 0.39 | **0.01** | + |
| | m638 | 0.39 | 0.45 | 0.18 | 0.34 | 0.36 | 0.32 | 0.31 | + |
| | m648 | 0.22 | 0.30 | 0.30 | 0.21 | 0.24 | 0.31 | 0.18 | - |
| | m652 | 0.32 | 0.25 | 0.14 | 0.45 | 0.38 | 0.41 | 0.42 | + |
| | m655 | 0.33 | 0.24 | 0.30 | 0.44 | 0.35 | 0.41 | 0.37 | + |
| | m657 | 0.09 | **0.04** | **0.04** | **0.01** | 0.23 | 0.23 | **0.03** | + |
| | m658 | 0.46 | 0.05 | 0.46 | 0.45 | 0.43 | 0.22 | 0.05 | + |
| AASenes | m659 | 0.40 | 0.46 | 0.46 | 0.32 | 0.21 | 0.33 | 0.37 | + |
| | m11 | 0.40 | 0.30 | 0.05 | 0.28 | 0.45 | 0.48 | 0.40 | + |
| | m229 | 0.26 | 0.33 | 0.32 | **0.03** | 0.27 | 0.32 | 0.31 | + |
| | m68 | -1.00 | -1.00 | 0.08 | 0.05 | **0.02** | **0.01** | **0.01** | - |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m533 | 0.35 | 0.47 | **0.03** | 0.48 | 0.34 | 0.47 | 0.40 | + |
| | m228 | 0.40 | 0.35 | 0.07 | 0.42 | 0.38 | 0.45 | 0.05 | + |
| | m226 | 0.29 | 0.33 | 0.19 | 0.49 | 0.49 | 0.46 | 0.14 | + |
| | m225 | **0.00** | **0.00** | -1.00 | **0.02** | **0.02** | -1.00 | 0.06 | - |
| | m32 | 0.40 | 0.17 | 0.42 | 0.23 | 0.23 | 0.48 | 0.45 | + |
| | m160 | **0.01** | 0.19 | 0.09 | 0.17 | 0.30 | 0.22 | 0.26 | + |
| | m179 | 0.07 | 0.15 | 0.11 | 0.13 | 0.07 | **0.02** | 0.23 | + |
| | m36 | 0.08 | 0.16 | 0.08 | 0.13 | **0.03** | 0.12 | **0.02** | - |
| | m472 | 0.19 | 0.32 | 0.20 | 0.34 | 0.28 | 0.30 | 0.09 | + |
| | m387 | 0.05 | 0.05 | -1.00 | **0.04** | 0.08 | **0.00** | **0.00** | |
| | m246 | 0.08 | **0.00** | -1.00 | **0.00** | **0.02** | 0.06 | 0.07 | - |
| | m491 | -1.00 | 0.22 | 0.08 | **0.00** | **0.02** | **0.04** | 0.08 | - |
| | m112 | 0.29 | 0.27 | 0.38 | 0.16 | 0.48 | 0.45 | **0.03** | + |
| | m498 | -1.00 | -1.00 | 0.15 | **0.04** | **0.02** | **0.03** | -1.00 | - |
| | m360 | 0.20 | 0.15 | 0.35 | 0.42 | 0.32 | 0.35 | 0.10 | + |
| | m361 | 0.05 | 0.05 | -1.00 | **0.02** | **0.01** | **0.01** | **0.00** | - |
| | m420 | 0.05 | **0.02** | 0.36 | 0.17 | 0.46 | 0.42 | 0.26 | + |
| | m323 | 0.12 | **0.03** | 0.34 | 0.21 | 0.14 | 0.13 | 0.07 | + |
| | m324 | 0.37 | 0.31 | 0.27 | 0.44 | 0.40 | 0.49 | 0.20 | + |
| | m512 | -1.00 | -1.00 | **0.04** | 0.05 | **0.04** | **0.04** | 0.12 | - |
| | m325 | 0.08 | **0.04** | -1.00 | 0.05 | **0.02** | 0.05 | 0.15 | - |
| | m320 | -1.00 | -1.00 | -1.00 | 0.10 | -1.00 | -1.00 | -1.00 | - |
| | m319 | 0.06 | 0.14 | 0.08 | 0.09 | 0.06 | 0.10 | 0.19 | - |
| | m326 | 0.30 | 0.47 | 0.10 | 0.37 | 0.24 | 0.45 | 0.39 | + |
| | m537 | 0.38 | 0.16 | **0.01** | 0.22 | 0.47 | 0.45 | 0.30 | + |
| | m321 | -1.00 | 0.05 | -1.00 | **0.02** | **0.02** | 0.05 | **0.03** | - |
| | m125 | 0.05 | 0.13 | 0.09 | 0.11 | 0.06 | **0.03** | 0.22 | - |
| | m483 | -1.00 | -1.00 | 0.04 | 0.03 | 0.03 | -1.00 | -1.00 | - |
| | m120 | 0.30 | 0.20 | 0.27 | 0.36 | 0.34 | 0.35 | 0.23 | + |
| | m429 | -1.00 | **0.04** | **0.04** | 0.06 | **0.02** | **0.01** | 0.00 | - |
| | m419 | -1.00 | -1.00 | **0.04** | **0.04** | **0.02** | -1.00 | -1.00 | - |
| | m252 | **0.00** | 0.09 | 0.11 | **0.04** | 0.34 | **0.04** | **0.02** | - |
| | m257 | -1.00 | -1.00 | -1.00 | **0.02** | 0.10 | -1.00 | -1.00 | - |
| | m476 | 0.29 | 0.19 | 0.12 | 0.39 | 0.41 | 0.39 | 0.26 | + |
| | m10 | -1.00 | -1.00 | -1.00 | **0.03** | **0.04** | **0.03** | 0.12 | - |
| Reldent | m251 | 0.42 | 0.31 | 0.41 | 0.35 | 0.09 | 0.37 | 0.05 | + |
| | m254 | -1.00 | -1.00 | -1.00 | **0.04** | **0.01** | 0.05 | -1.00 | - |
| | m256 | 0.17 | 0.06 | 0.28 | 0.30 | 0.20 | 0.23 | 0.44 | + |
| | m470 | 0.44 | 0.42 | 0.11 | 0.36 | 0.29 | 0.38 | **0.01** | + |
| | m116 | 0.05 | **0.03** | 0.07 | 0.16 | 0.08 | 0.08 | **0.03** | + |
| | m253 | 0.21 | 0.29 | 0.29 | 0.12 | 0.22 | 0.31 | 0.13 | - |
| MelE | m432 | -1.00 | -1.00 | **0.04** | **0.03** | **0.02** | **0.04** | **0.00** | - |
| | m431 | 0.34 | 0.44 | 0.31 | 0.24 | 0.47 | 0.10 | **0.04** | + |
| | m255 | 0.47 | 0.29 | 0.22 | 0.20 | 0.44 | 0.27 | 0.10 | + |
| | m200 | 0.19 | 0.17 | 0.07 | 0.49 | 0.30 | 0.33 | **0.03** | + |
| | m195 | 0.31 | 0.20 | 0.31 | 0.44 | 0.46 | 0.42 | 0.27 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m194 | 0.31 | 0.23 | 0.17 | 0.16 | 0.30 | 0.06 | 0.21 | + |
| | m12 | 0.18 | 0.42 | 0.01 | 0.31 | 0.26 | 0.30 | 0.11 | + |
| | m196 | 0.42 | 0.09 | 0.09 | 0.10 | 0.06 | 0.06 | 0.06 | - |
| | m192 | 0.35 | 0.27 | **0.04** | **0.02** | 0.48 | 0.30 | 0.47 | + |
| AASens | m508 | 0.15 | 0.07 | 0.42 | 0.31 | 0.41 | 0.39 | 0.40 | + |
| | m150 | -1.00 | -1.00 | -1.00 | **0.03** | -1.00 | -1.00 | -1.00 | - |
| Eye | m467 | 0.45 | 0.42 | 0.09 | 0.38 | 0.18 | 0.38 | 0.29 | + |
| | m265 | 0.07 | 0.04 | 0.33 | 0.19 | 0.13 | 0.15 | **0.01** | + |
| COND | m224 | 0.05 | 0.09 | 0.09 | 0.07 | 0.05 | 0.09 | 0.21 | - |
| | m219 | 0.17 | **0.01** | 0.25 | 0.13 | **0.04** | 0.06 | **0.04** | - |
| | m221 | 0.16 | **0.02** | 0.07 | 0.06 | 0.09 | 0.06 | 0.30 | + |
| Wtloss | m222 | -1.00 | -1.00 | -1.00 | **0.00** | **0.03** | -1.00 | -1.00 | - |
| | m708 | 0.27 | 0.13 | 0.23 | 0.39 | 0.47 | 0.35 | 0.22 | + |
| | m709 | 0.27 | 0.17 | 0.42 | 0.39 | 0.15 | 0.35 | 0.23 | + |
| | m457 | **0.04** | 0.12 | **0.02** | 0.15 | 0.30 | **0.02** | 0.08 | - |
| | m223 | 0.40 | 0.33 | 0.21 | 0.43 | 0.32 | 0.45 | 0.23 | + |
| | m220 | 0.38 | 0.30 | 0.32 | 0.46 | 0.45 | **0.02** | 0.16 | + |
| | m688 | 0.26 | 0.22 | 0.30 | 0.40 | 0.09 | 0.31 | 0.14 | + |
| | m696 | 0.46 | 0.43 | 0.20 | 0.47 | 0.47 | 0.47 | 0.43 | + |
| | m686 | 0.45 | 0.46 | 0.25 | 0.44 | 0.45 | 0.43 | 0.35 | + |
| | m690 | 0.45 | 0.38 | 0.25 | 0.15 | 0.45 | 0.49 | 0.37 | + |
| | m689 | 0.44 | 0.34 | 0.45 | 0.40 | **0.02** | 0.28 | 0.06 | + |
| | m695 | 0.18 | 0.05 | 0.36 | 0.37 | 0.47 | 0.44 | 0.35 | + |
| Eye | m693 | 0.25 | 0.39 | 0.12 | 0.47 | 0.37 | 0.38 | 0.21 | + |
| Eye | m694 | 0.17 | 0.08 | 0.15 | 0.35 | 0.49 | 0.43 | 0.34 | + |
| COND | m243 | -1.00 | -1.00 | -1.00 | 0.05 | 0.14 | -1.00 | -1.00 | - |
| | m212 | -1.00 | -1.00 | -1.00 | 0.07 | -1.00 | -1.00 | -1.00 | - |
| | m216 | 0.20 | 0.12 | 0.15 | 0.34 | 0.10 | 0.18 | **0.03** | + |
| | m259 | 0.32 | 0.47 | 0.23 | 0.42 | 0.37 | 0.40 | 0.34 | + |
| | m261 | 0.06 | **0.03** | 0.14 | 0.12 | 0.12 | 0.10 | **0.03** | + |
| | m260 | 0.29 | 0.38 | 0.16 | 0.48 | 0.43 | 0.45 | 0.11 | + |
| | m94 | **0.00** | **0.01** | **0.00** | **0.00** | **0.03** | **0.04** | **0.00** | - |
| | m505 | -1.00 | -1.00 | **0.04** | 0.07 | **0.05** | -1.00 | -1.00 | - |
| | m443 | -1.00 | -1.00 | -1.00 | **0.01** | **0.02** | **0.00** | **0.02** | - |
| | m487 | -1.00 | -1.00 | -1.00 | **0.02** | **0.04** | **0.02** | **0.01** | - |
| | m230 | -1.00 | -1.00 | -1.00 | 0.09 | -1.00 | -1.00 | -1.00 | - |
| WTLoss | m232 | 0.19 | 0.17 | **0.03** | 0.32 | 0.33 | 0.28 | 0.32 | + |
| | m235 | -1.00 | **0.04** | -1.00 | **0.03** | **0.02** | **0.01** | 0.06 | - |
| | m233 | **0.03** | 0.27 | 0.23 | 0.12 | 0.07 | **0.04** | **0.00** | - |
| COND | m231 | 0.05 | 0.09 | 0.09 | 0.05 | 0.06 | 0.06 | 0.22 | - |
| LEN | m479 | -1.00 | -1.00 | **0.04** | **0.00** | **0.02** | **0.00** | **0.00** | - |
| | m237 | 0.06 | 0.08 | 0.11 | 0.05 | **0.03** | 0.05 | 0.25 | - |
| MEl E | m239 | -1.00 | **0.04** | 0.15 | 0.13 | **0.03** | 0.11 | 0.13 | - |
| MEL A | m500 | 0.09 | 0.06 | 0.13 | **0.01** | 0.07 | 0.06 | 0.06 | - |
| | m238 | 0.40 | 0.37 | 0.24 | 0.39 | 0.32 | 0.39 | 0.23 | + |
| | m503 | 0.28 | 0.31 | 0.06 | 0.13 | 0.04 | 0.28 | 0.09 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m397 | 0.17 | 0.35 | 0.33 | 0.48 | 0.36 | 0.47 | 0.10 | + |
| Eye | m161 | -1.00 | -1.00 | -1.00 | 2.00 | -1.00 | -1.00 | -1.00 | - |
| | m410 | 0.28 | **0.02** | 0.39 | 0.25 | 0.20 | **0.01** | **0.02** | + |
| | m206 | 0.14 | 0.12 | 0.03 | 0.49 | 0.47 | 0.32 | 0.21 | + |
| | m205 | 0.07 | 0.15 | 0.11 | 0.09 | 0.24 | **0.03** | 0.24 | + |
| | m210 | 0.32 | 0.41 | 0.38 | 0.06 | 0.36 | **0.02** | 0.41 | + |
| | m204 | **0.00** | **0.01** | -1.00 | **0.01** | 0.15 | -1.00 | 0.06 | - |
| Mel E | m209 | 0.10 | 0.05 | 0.07 | 0.32 | 0.09 | 0.21 | 0.07 | + |
| | m207 | 0.12 | 0.38 | 0.05 | 0.28 | 0.31 | 0.29 | **0.04** | + |
| | m494 | -1.00 | -1.00 | 0.18 | **0.00** | 0.19 | -1.00 | -1.00 | - |
| | m511 | 0.27 | 0.17 | 0.06 | 0.40 | 0.47 | 0.38 | 0.38 | + |
| | m241 | 0.05 | **0.04** | -1.00 | **0.04** | **0.02** | **0.04** | **0.01** | - |
| | m242 | 0.25 | 0.32 | 0.07 | 0.12 | 0.26 | 0.31 | 0.42 | + |
| | m175 | 0.46 | 0.14 | 0.34 | **0.03** | 0.47 | 0.47 | 0.09 | + |
| | m77 | 0.27 | 0.33 | 0.23 | 0.26 | **0.03** | 0.25 | 0.26 | + |
| | m434 | 0.39 | 0.44 | **0.03** | 0.37 | 0.38 | 0.36 | **0.02** | + |
| | m61 | 0.27 | 0.18 | 0.33 | 0.46 | 0.46 | 0.42 | 0.14 | + |
| | m527 | 0.46 | 0.40 | 0.09 | 0.48 | 0.20 | 0.31 | 0.07 | + |
| | m295 | 0.24 | 0.18 | 0.05 | 0.33 | 0.48 | 0.34 | 0.37 | + |
| | m382 | 0.06 | 0.14 | **0.04** | **0.04** | 0.28 | 0.11 | 0.24 | - |
| | m186 | 0.41 | 0.25 | 0.06 | 0.50 | 0.26 | 0.41 | 0.41 | + |
| | m7 | -1.00 | -1.00 | -1.00 | **0.01** | 0.07 | 0.24 | **0.01** | - |
| | m504 | 0.29 | 0.05 | 0.10 | 0.22 | 0.47 | 0.18 | 0.07 | + |
| | m456 | 0.07 | 0.36 | 0.25 | 0.09 | 0.09 | 0.17 | **0.02** | + |
| | m304 | 0.40 | 0.27 | 0.37 | 0.29 | 0.25 | 0.10 | 0.40 | + |
| | m299 | -1.00 | -1.00 | -1.00 | 0.13 | **0.03** | -1.00 | -1.00 | - |
| | m3 | 0.35 | 0.16 | 0.41 | 0.48 | 0.47 | 0.34 | 0.25 | + |
| | m438 | 0.34 | 0.25 | 0.42 | 0.41 | 0.42 | 0.39 | 0.47 | + |
| | m471 | 0.22 | 0.06 | 0.07 | 0.37 | 0.20 | 0.32 | 0.40 | + |
| | m322 | 0.15 | 0.25 | 0.10 | 0.37 | 0.33 | 0.29 | 0.11 | + |
| | m203 | -1.00 | -1.00 | -1.00 | **0.01** | -1.00 | -1.00 | -1.00 | - |
| | m208 | 0.48 | 0.07 | 0.44 | 0.43 | 0.46 | 0.43 | 0.43 | + |
| | m351 | -1.00 | -1.00 | -1.00 | **0.00** | 0.14 | -1.00 | -1.00 | - |
| | m510 | 0.22 | **0.02** | 0.24 | **0.01** | 0.22 | 0.23 | **0.02** | + |
| | m146 | 0.31 | 0.21 | 0.18 | 0.44 | 0.47 | 0.48 | 0.23 | + |
| | m448 | 0.27 | 0.17 | 0.22 | 0.31 | 0.45 | 0.41 | **0.02** | + |
| | m480 | **0.00** | 0.07 | **0.04** | **0.00** | -1.00 | -1.00 | -1.00 | - |
| | m22 | 0.48 | 0.42 | 0.05 | 0.47 | 0.48 | 0.30 | 0.14 | + |
| | m144 | 0.33 | 0.45 | 0.40 | 0.44 | 0.37 | 0.32 | 0.33 | + |
| | m214 | 0.28 | 0.34 | 0.20 | 0.39 | 0.47 | 0.34 | 0.47 | + |
| | m217 | 0.16 | 0.20 | 0.09 | 0.32 | 0.24 | 0.26 | 0.30 | + |
| | m490 | -1.00 | -1.00 | -1.00 | **0.04** | -1.00 | -1.00 | -1.00 | - |
| | m128 | 0.27 | 0.14 | 0.24 | 0.40 | 0.15 | 0.46 | 0.10 | + |
| | m458 | 0.28 | 0.32 | 0.42 | 0.50 | 0.39 | 0.44 | 0.40 | - |
| | m524 | 0.46 | 0.31 | 0.27 | **0.00** | 0.31 | 0.40 | 0.32 | + |
| | m385 | 0.26 | 0.19 | **0.00** | 0.43 | 0.36 | 0.41 | 0.18 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m356 | 0.32 | 0.22 | 0.36 | 0.23 | 0.35 | 0.37 | 0.24 | + |
| | m367 | 0.42 | 0.47 | 0.41 | 0.47 | 0.42 | 0.33 | 0.36 | - |
| | m374 | 0.06 | 0.08 | 0.11 | 0.26 | 0.06 | **0.03** | **0.02** | + |
| | m375 | 0.05 | 0.13 | 0.09 | 0.19 | 0.05 | 0.13 | 0.21 | - |
| | m378 | 0.43 | 0.43 | 0.07 | 0.27 | 0.24 | 0.43 | 0.42 | + |
| | m393 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | 0.18 | 0.06 | - |
| | m394 | 0.22 | 0.09 | 0.25 | 0.15 | **0.04** | 0.06 | 0.36 | - |
| | m395 | -1.00 | -1.00 | -1.00 | -1.00 | 0.18 | -1.00 | -1.00 | - |
| | m399 | -1.00 | **0.04** | -1.00 | **0.02** | **0.02** | 0.08 | -1.00 | - |
| | m405 | 0.41 | 0.32 | 0.12 | 0.47 | 0.28 | 0.47 | 0.06 | + |
| | m406 | 0.19 | 0.10 | 0.16 | 0.50 | 0.29 | 0.30 | 0.20 | + |
| | m409 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | 0.14 | -1.00 | - |
| | m412 | -1.00 | -1.00 | -1.00 | **0.02** | 0.05 | -1.00 | -1.00 | - |
| | m415 | 0.25 | 0.43 | 0.42 | 0.41 | 0.28 | 0.07 | **0.04** | + |
| | m426 | -1.00 | -1.00 | 0.07 | **0.01** | 0.05 | -1.00 | -1.00 | - |
| | m435 | 0.39 | 0.43 | 0.28 | 0.41 | 0.47 | 0.46 | 0.41 | + |
| | m437 | 0.21 | 0.19 | 0.10 | 0.49 | 0.33 | 0.34 | 0.23 | + |
| | m441 | 0.47 | 0.43 | 0.33 | 0.48 | 0.47 | 0.48 | 0.45 | + |
| | m445 | 0.33 | 0.25 | 0.34 | 0.46 | 0.47 | 0.43 | 0.42 | + |
| | m450 | 0.32 | 0.37 | **0.03** | 0.37 | 0.33 | **0.02** | 0.08 | + |
| | m455 | 0.35 | 0.42 | 0.36 | 0.31 | 0.33 | 0.31 | 0.37 | + |
| | m462 | 0.48 | 0.42 | 0.42 | 0.45 | 0.10 | 0.44 | 0.45 | + |
| | m464 | 0.35 | **0.02** | 0.21 | 0.15 | 0.35 | 0.34 | **0.03** | - |
| | m465 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | 0.05 | -1.00 | - |
| | m469 | 0.05 | 0.09 | 0.31 | 0.15 | 0.07 | **0.04** | **0.01** | - |
| | m475 | -1.00 | -1.00 | **0.04** | 0.07 | 0.17 | -1.00 | -1.00 | - |
| | m478 | 0.34 | 0.14 | 0.07 | 0.30 | 0.33 | **0.01** | 0.05 | + |
| | m484 | 0.32 | 0.27 | **0.02** | 0.49 | **0.01** | 0.39 | 0.30 | + |
| | m496 | -1.00 | -1.00 | -1.00 | 0.04 | 0.06 | -1.00 | -1.00 | - |
| | m509 | 0.46 | 0.40 | 0.06 | 0.46 | 0.49 | 0.46 | 0.31 | + |
| | m514 | 0.34 | 0.24 | 0.08 | 0.34 | 0.48 | 0.29 | 0.17 | + |
| | m515 | 0.41 | 0.14 | 0.06 | 0.21 | 0.37 | 0.34 | 0.35 | + |
| | m522 | 0.05 | 0.25 | 0.25 | 0.32 | 0.21 | 0.22 | 0.04 | + |
| | m526 | -1.00 | -1.00 | 0.15 | 0.08 | 0.06 | 0.06 | 0.06 | - |
| | m528 | 0.46 | 0.04 | 0.44 | 0.38 | 0.44 | 0.40 | 0.25 | + |
| | m531 | **0.03** | 0.06 | 0.15 | **0.02** | 0.33 | 0.08 | 0.06 | + |
| | m532 | 0.23 | 0.27 | 0.05 | 0.39 | 0.37 | 0.38 | 0.06 | + |
| | m714 | 0.21 | 0.45 | 0.24 | 0.49 | 0.27 | 0.47 | 0.45 | + |
| | m716 | -1.00 | -1.00 | -1.00 | 0.05 | **0.02** | **0.03** | 0.05 | - |
| | m717 | 0.46 | 0.46 | 0.27 | 0.43 | 0.33 | 0.30 | **0.02** | + |
| | m719 | 0.09 | 0.08 | 0.22 | 0.13 | 0.43 | 0.44 | 0.12 | + |
| | m720 | 0.24 | 0.30 | 0.19 | 0.43 | 0.44 | 0.13 | 0.25 | + |
| | m721 | -1.00 | -1.00 | **0.04** | **0.02** | **0.02** | **0.04** | **0.01** | - |
| | m722 | -1.00 | -1.00 | 0.05 | 0.07 | **0.02** | **0.03** | **0.00** | - |
| | m723 | 0.21 | 0.27 | 0.40 | **0.02** | 0.04 | 0.06 | 0.05 | - |
| | m724 | 0.33 | 0.22 | 0.38 | 0.43 | 0.34 | 0.46 | 0.19 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m267 | -1.00 | -1.00 | -1.00 | **0.00** | -1.00 | -1.00 | -1.00 | - |
| | m436 | 0.46 | **0.03** | 0.07 | 0.38 | 0.26 | 0.40 | 0.36 | + |
| | m121 | 0.13 | 0.05 | **0.04** | 0.24 | 0.21 | 0.21 | **0.04** | + |
| | m264 | 0.06 | 0.08 | 0.11 | 0.07 | 0.21 | **0.03** | 0.05 | - |
| | m244 | 0.43 | 0.29 | 0.21 | 0.27 | 0.39 | 0.50 | 0.44 | + |
| | m245 | 0.42 | 0.29 | 0.24 | 0.34 | 0.42 | 0.49 | 0.49 | + |
| | m1 | -1.00 | -1.00 | -1.00 | 0.07 | 0.07 | 0.08 | -1.00 | - |
| | m250 | 0.29 | 0.22 | 0.22 | 0.39 | 0.41 | 0.41 | 0.28 | + |
| | m5 | -1.00 | 0.04 | 0.06 | -1.00 | -1.00 | -1.00 | 0.07 | - |
| | m9 | 0.31 | 0.22 | 0.20 | 0.38 | 0.09 | 0.23 | 0.46 | + |
| | m20 | 0.26 | 0.18 | 0.31 | 0.39 | 0.26 | 0.30 | 0.07 | - |
| | m23 | 0.13 | 0.07 | 0.26 | 0.39 | 0.21 | 0.21 | **0.04** | + |
| | m29 | 0.21 | 0.33 | 0.29 | 0.41 | 0.12 | 0.31 | **0.03** | - |
| | m42 | 0.21 | 0.13 | 0.12 | 0.36 | 0.40 | 0.32 | 0.13 | + |
| | m44 | 0.23 | 0.15 | **0.01** | 0.11 | 0.23 | 0.17 | 0.08 | - |
| | m45 | 0.12 | 0.08 | 0.23 | 0.33 | 0.29 | 0.21 | 0.29 | + |
| | m59 | 0.26 | 0.15 | 0.09 | 0.40 | 0.43 | 0.40 | 0.36 | + |
| | m85 | 0.24 | 0.30 | 0.30 | 0.24 | 0.24 | 0.31 | 0.18 | + |
| | m89 | 0.08 | 0.41 | 0.14 | 0.38 | 0.41 | 0.39 | 0.37 | + |
| | m92 | -1.00 | -1.00 | -1.00 | 0.09 | **0.04** | -1.00 | -1.00 | - |
| | m97 | -1.00 | -1.00 | -1.00 | 0.09 | -1.00 | 0.11 | 0.14 | - |
| | m100 | 0.10 | 0.05 | 0.23 | 0.39 | 0.40 | 0.48 | 0.20 | + |
| | m104 | 0.28 | 0.46 | 0.45 | 0.16 | 0.34 | 0.37 | 0.21 | + |
| | m113 | 0.22 | 0.10 | 0.12 | 0.43 | 0.33 | 0.23 | 0.44 | + |
| | m122 | 0.44 | **0.02** | **0.02** | 0.20 | 0.40 | 0.38 | **0.01** | + |
| | m124 | 0.32 | 0.08 | 0.38 | 0.45 | 0.32 | 0.35 | 0.41 | + |
| | m126 | 0.30 | **0.02** | 0.37 | 0.45 | 0.45 | 0.41 | 0.05 | + |
| | m130 | 0.44 | 0.45 | 0.38 | **0.02** | 0.44 | 0.45 | 0.18 | + |
| | m135 | 0.38 | 0.27 | 0.06 | 0.30 | 0.29 | 0.29 | 0.27 | + |
| | m136 | 0.22 | 0.26 | 0.09 | 0.25 | **0.03** | **0.04** | **0.02** | + |
| | m137 | -1.00 | -1.00 | 0.16 | **0.01** | -1.00 | -1.00 | -1.00 | - |
| | m138 | -1.00 | -1.00 | -1.00 | 0.05 | **0.01** | **0.00** | **0.01** | - |
| | m139 | 0.21 | 0.25 | **0.01** | 0.20 | 0.15 | 0.18 | 0.05 | + |
| | m141 | 0.26 | 0.16 | 0.21 | 0.46 | 0.39 | 0.22 | 0.17 | + |
| | m142 | 0.42 | 0.47 | 0.42 | 0.46 | 0.42 | 0.48 | 0.14 | + |
| | m143 | 0.22 | 0.27 | 0.25 | 0.31 | 0.22 | 0.24 | 0.35 | - |
| | m711 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | **0.02** | -1.00 | - |
| | m712 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | 0.08 | -1.00 | - |
| | m148 | **0.04** | **0.02** | -1.00 | 0.08 | -1.00 | 0.15 | -1.00 | - |
| | m151 | 0.25 | 0.32 | 0.29 | 0.38 | 0.13 | 0.46 | 0.40 | - |
| | m152 | 0.22 | 0.21 | 0.10 | 0.35 | 0.38 | 0.45 | 0.09 | + |
| | m263 | 0.07 | 0.15 | 0.11 | 0.06 | 0.33 | 0.13 | 0.25 | - |
| | m278 | 0.08 | 0.28 | 0.41 | 0.46 | 0.25 | 0.10 | **0.02** | + |
| | m300 | 0.44 | 0.38 | 0.27 | 0.44 | 0.37 | 0.45 | 0.40 | + |
| | m332 | 0.09 | 0.05 | -1.00 | 0.09 | 0.03 | 0.14 | **0.04** | - |
| | m352 | 0.34 | 0.19 | 0.13 | 0.46 | 0.45 | 0.47 | 0.18 | + |

| QTL | M | N1 | N2 | N3 | O1 | O2O3 | O4O6 | O8 | P |
|---|---|---|---|---|---|---|---|---|---|
| | m725 | -1.00 | -1.00 | 0.08 | 0.07 | **0.02** | **0.00** | **0.00** | - |
| | m727 | 0.44 | 0.16 | 0.06 | 0.44 | 0.09 | 0.24 | 0.18 | + |
| | m728 | 0.24 | 0.30 | 0.05 | **0.04** | 0.24 | 0.31 | 0.13 | - |
| | m729 | -1.00 | -1.00 | **0.00** | **0.04** | **0.02** | **0.00** | **0.00** | - |
| | m730 | 0.07 | **0.03** | 0.04 | 0.21 | 0.38 | 0.44 | 0.35 | + |
| | m731 | 0.17 | 0.09 | **0.03** | 0.33 | 0.22 | 0.31 | 0.18 | + |
| | m732 | 0.28 | 0.27 | 0.21 | 0.45 | 0.44 | 0.49 | 0.23 | + |
| | m734 | 0.22 | 0.27 | 0.21 | **0.04** | 0.05 | 0.06 | 0.10 | - |
| | m735 | 0.06 | **0.04** | **0.01** | 0.15 | 0.10 | 0.12 | 0.05 | + |
| | m736 | 0.24 | 0.28 | 0.05 | 0.23 | 0.07 | 0.23 | 0.20 | + |
| | m737 | -1.00 | -1.00 | **0.04** | **0.04** | **0.01** | -1.00 | -1.00 | - |
| | m738 | 0.14 | 0.17 | **0.04** | 0.08 | 0.13 | 0.16 | **0.02** | + |
| | m739 | -1.00 | -1.00 | **0.04** | 0.07 | 0.17 | -1.00 | -1.00 | - |
| | m740 | 0.22 | 0.27 | 0.16 | 0.09 | 0.22 | 0.26 | 0.35 | - |

Table S3.2. Summary of all $F_{ST}$ values from the markers in the study assigned to the QTL locus. Column P represents assignment of each marker based on its presence in “surface SNPs” (MAF>5%) or “cave SNPs” (MAF<5%) and it is labeled as + or –, respectively. -1 value determines that the variance was not possible to calculate since there was no variation between particular population and the surface population. Values in the table represent P-values <0.05 as identified by coalescent simulations in ARELQUIN. Red labels represents loci that are detected as significant outliers across old and new caves in at least three populations, while blue labels represent markers that are significant in at least three populations in only new or old group. Gray blocks represent the block of QTL locus.

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100
Fax (+351) 214 411 277

**www.itqb.unl.pt**